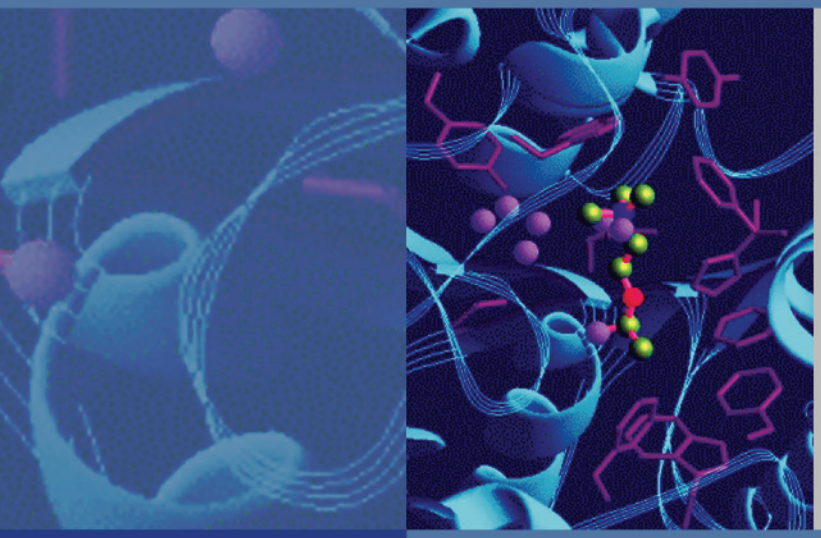


Xcalibur

SIEVE

User Guide

XCALI-97186 Revision B July 2007

For Research Use Only
Not for use in Diagnostic Procedures

Thermo SCIENTIFIC

© 2007 Thermo Fisher Scientific Inc. All rights reserved.

Microsoft® Windows® XP is a registered trademark of Microsoft Corporation in the United States and other countries.

SIEVE™, ChromAlign™, LTQ™ XL, LTQ Orbitrap™, LTQ FT Ultra™, and BioWorks™ are trademarks and Xcalibur® is a registered trademark of Thermo Fisher Scientific Inc.

DecisionSite™ is a trademark and Spotfire® is a registered trademark of Spotfire, Inc.

SEQUEST® is a registered trademark of the University of Washington.

Thermo Fisher Scientific Inc. provides this document to its customers with a product purchase to use in the product operation. This document is copyright protected and any reproduction of the whole or any part of this document is strictly prohibited, except with the written authorization of Thermo Fisher Scientific Inc.

The contents of this document are subject to change without notice. All technical information in this document is for reference purposes only. System configurations and specifications in this document supersede all previous information received by the purchaser.

Thermo Fisher Scientific Inc. makes no representations that this document is complete, accurate or error-free and assumes no responsibility and will not be liable for any errors, omissions, damage or loss that might result from any use of this document, even if the information in the document is followed properly.

This document is not part of any sales contract between Thermo Fisher Scientific Inc. and a purchaser. This document shall in no way govern or modify any Terms and Conditions of Sale, which Terms and Conditions of Sale shall govern all conflicting information between the two documents.

Release history:

For Research Use Only. Not regulated for medical or veterinary diagnostic use by U.S. FDA or other competent authorities.

Contents

Preface

SIEVE User Guide

This guide describes how to use SIEVE to compare multiple LC/MS data sets to determine the statistically significant differences between samples and identify the corresponding peptides and proteins.

To help you make the most of SIEVE, review the following topics:

- Viewing the Help
- System Requirements
- Getting a License
- Safety and Special NoticesContacting Us
- Contacting Us

Viewing the Help

In addition to this guide, Thermo Scientific provides Help that you can access from the software.

System Requirements

The minimum hardware requirements for SIEVE 1.1.0 are as follows:

- 2 GHz processor
- 1 GB RAM
- 40 GB hard disk (formatted with NTFS)
- CD-ROM drive
- Display resolution 1280 x 1024 (XGA with small/standard fonts)

To install and use SIEVE 1.1.0 requires the following software:

- Windows® XP with SP2
- Microsoft® Office XP Professional SP2
- BioWorks™ 3.3.1
- Xcalibur® 2.0.5 or higher
- Microsoft® Internet Explorer 6.0

Getting a License

❖ **To get a license for SIEVE**

1. Install the SIEVE software according to the instructions on the CD cover.
2. Get a new license code from Thermo Fisher Scientific.
3. Install the new license code.

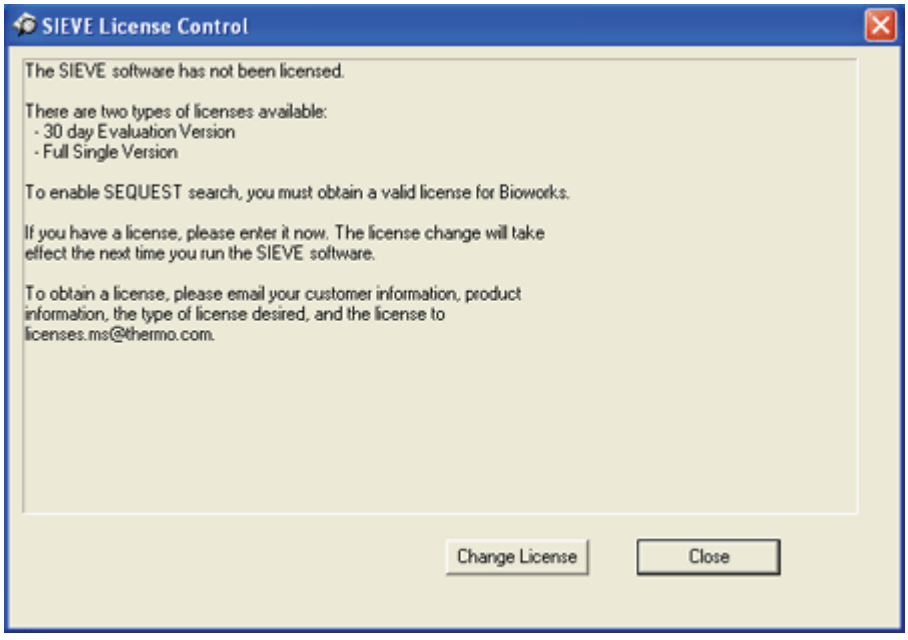
Getting a New License Code

The following instructions describe using e-mail to get a license code. If you do not have access to e-mail, you can get a license by fax.

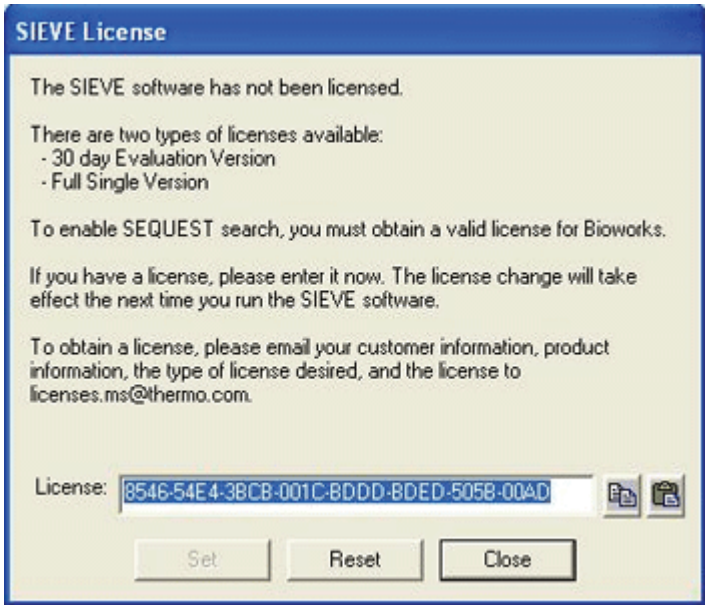
Note Do not click Reset unless instructed to do so by an authorized Thermo Fisher Scientific representative.

❖ **To get a license code**

1. Install the SIEVE software by clicking the **SIEVE** install icon.
2. Follow the prompts to install the software.
3. To start SIEVE, choose **Start > All Programs > Xcalibur > SIEVE** or click the **SIEVE** icon on your computer desktop.
   If you do not have a current license, the following screen appears:



4. Click **Change License** to display the license dialog box:



Note Do not click Reset unless instructed to do so by an authorized Thermo Fisher Scientific representative.

5. Highlight the license key in the License text box.
6. Press CTRL+C to copy the key to the Windows clipboard.
7. Send an e-mail message to licenses.ms@thermo.com:
   a. Type **license request** in the subject line of the e-mail, and paste the license key into the body of the e-mail.
   b. Provide your customer information.
   c. Locate the bar code on the back of the SIEVE CD jewel case. Type the serial number that appears below the bar code into the body of the e-mail.

Thermo Fisher Scientific Customer Support sends you a new license code. To install the new license code, go to the next section.

Installing a New License Code

Once you have received your new license code, install it as follows:

❖ **To install the license code**

1. To start SIEVE, choose **Start > All Programs > Xcalibur > SIEVE** or click the **SIEVE** icon on your computer desktop.
2. Click **Change License** to display the SIEVE License dialog box.

   The SIEVE License dialog box appears.
3. In the License text box, type the new license number and click **Set**.
4. To accept the change in license, click **Yes**.
5. To close the SIEVE Browser License dialog, click **Close**.

Note Do not choose Reset at this point or the license software invalidates the license you just received.

Safety and Special Notices

Make sure you follow the precautionary statements presented in this guide. The safety and other special notices appear in boxes.

Safety and special notices include the following:

CAUTION Highlights hazards to humans, property, or the environment. Each CAUTION notice is accompanied by an appropriate CAUTION symbol.

IMPORTANT Highlights information necessary to prevent damage to software, loss of data, or invalid test results; or may contain information that is critical for optimal performance of the system.

Note Highlights information of general interest.

Tip Helpful information that can make a task easier.

Contacting Us

There are several ways to contact Thermo Fisher Scientific.

❖ To contact Technical Support

Phone	800-685-9535
Fax	561-688-8736
E-mail	TechSupport.C+MS@thermofisher.com
Knowledge base	www.thermokb.com

Find software updates and utilities to download at www.mssupport.thermo.com.

❖ To order products or to get product updates

Phone	800-532-4752
Fax	561-688-8731
Web site	www.thermo.com/finnigan

❖ To suggest changes to documentation or to Help

- Fill out a reader survey online at www.thermo.com/lcms-techpubs.
- Send an e-mail message to the Technical Publications Editor at techpubs.finnigan-lcms@thermofisher.com.

Overview to SIEVE

SIEVE software performs semi-quantitative differential analyses of sample populations by comparing spectral information from LC/MS analyses of control (healthy) and disease (treatment) samples. Any changes between the two sample sets indicate a differential protein expression. Elements from raw files that show significant statistical differences are sent for a SEQUEST ® database search to determine peptide and protein identification.

The following sections provide a high-level overview of SIEVE:

Understanding SIEVE

SIEVE is an automated software package for the label-free, semi-quantitative differential expression of proteins and peptides. It performs comparative analyses of sample populations by comparing the raw spectral information from LC/MS analyses of control (healthy) and disease (treatment) samples to determine if there are changes among the two sample sets which may indicate differential protein expression.

SIEVE provides a statistically rigorous tool for analyzing the data acquired in protein biomarker discovery experiments. SIEVE can analyze as many as 100 LC/MS data files in an experiment that compares 50 sample files to 50 control files in a single operation. It can also compare a minimum of four files (select two control files and two sample files so that the pValue is valid).

SIEVE employs a Chromalign™ algorithm that allows comparison of similar chromatographic surfaces. From here, it uses a proprietary iterative process called Recursive Base Peak Framing to find statistically meaningful differences. For more information about the ChromAlign procedure, see Appendix A, "Understanding the ChromAlign Process." For more information about the framing process, see Appendix B, "Understanding the Framing Process."

To identify proteins, SIEVE performs a comparative analysis prior to a SEQUEST database search. To identify peptides and proteins, send elements from the raw data files that show significant statistical differences for a SEQUEST database search. SIEVE maximizes computational time and resources by providing the option to search only those features that have shown statistically significant change, accelerating the process of finding potential biomarkers.

SIEVE uses the MS intensities from the raw LC/MS data without the need to manipulate or model peaks to find statistical differences. The process is label-free, using no isotopic tags or labels of any kind.

SIEVE calculates a pValue to indicate the statistical significance of the expression ratio of each putative biomarker, providing more confidence in the results. SIEVE displays intermediate and final results using the Spotfire™ DecisionSite® software, a powerful interactive visual environment for data review.

Use SIEVE to identify protein biomarkers that indicate the presence of disease or identify potential targets in the drug discovery and development process.

About the SIEVE Procedure

The following section describes the SIEVE procedure. For specific information on each step, see Chapter 2, "Setting up a Procedure."

Step 1: Acquire raw data

Discovery experiments for global biomarker often consist of thousands of complex MS and MS^n data. SIEVE accommodates both high-resolution and low-resolution data, including data from the LTQ™ XL linear ion trap mass spectrometer, as well as the hybrid LTQ Orbitrap™ and LTQ FT Ultra™ instruments.

Step 2: Combine multiple data

An effective biomarker experiment requires the comparison of at least two raw data files, but to be statistically meaningful, the experiment must include several control samples compared to several treatment (disease) samples in the form of technical replicates and bio-replicates. SIEVE supports comparing 50 chromatograms to another 50 chromatograms, totalling 100 raw data files. To facilitate this comparison, SIEVE presents the data in one consolidated view that you can quickly view for differences. SIEVE can combine and compare from 2 to 100 data files.

Step 3: Align ... Chromatograms

Chromatographic alignment is an important step in the SIEVE pipeline since all the experimental data (up to 100 data files) is analyzed in the common, multivariate space described in Step 2. Consequently, SIEVE must correct for any chromatographic variation across the samples under analysis. For this purpose, SIEVE employs a unique chromatographic alignment algorithm called ChromAlign.

Step 4: Frame the data

SIEVE combines each raw file on a 3D plane of *m/z* versus retention time versus intensity.

SIEVE automates the process of determining differential expression by employing a proprietary algorithm called Recursive Base Peak Framing. This algorithm statistically compares the intensity of the peaks within each frame—areas in the *m/z* by chromatographic-time plane where differential expression is likely to have occurred.

The primary plot generated by SIEVE within Spotfire displays retention time versus *m/z*. Each blue rectangle represents a single frame. The consolidated plot is made up of several hundred to several thousand frames, the darker of which indicate a more statistically significant difference.

Step 5: Analyze the data using SEQUEST to identify proteins

Select SEQUEST to perform a database search only on frames determined to have confident pValues. This step identifies peptides and proteins that might represent differential expression. Using SIEVE to pre-filter data greatly reduces the number of spectra that need to be searched, significantly decreases the time spent identifying proteins, and increases the throughput of complex sample sets.

Step 6: Review Interactive Results

SIEVE provides interactive results using several different Spotfire plots, specifically chosen to maximize the information extracted from these types of biomarker discovery experiments. Breakpoints after ChromAlign, Recursive Base Peak Framing, and SEQUEST provide the option of checking intermediate results before proceeding or adjusting parameters.

Setting up a Procedure

The following sections describe setting up and running a procedure for SIEVE using two sets of raw files:

- Starting a New Experiment
- Setting Parameters
- Analyzing Files
- Framing the Data
- Identifying Proteins and Peptides
- Viewing a Saved Experiment

Starting a New Experiment

❖ To start a new experiment and select data for the analysis

1. To start SIEVE, choose **Start > All Programs > Xcalibur > SIEVE** or click the **SIEVE** icon on your computer desktop.

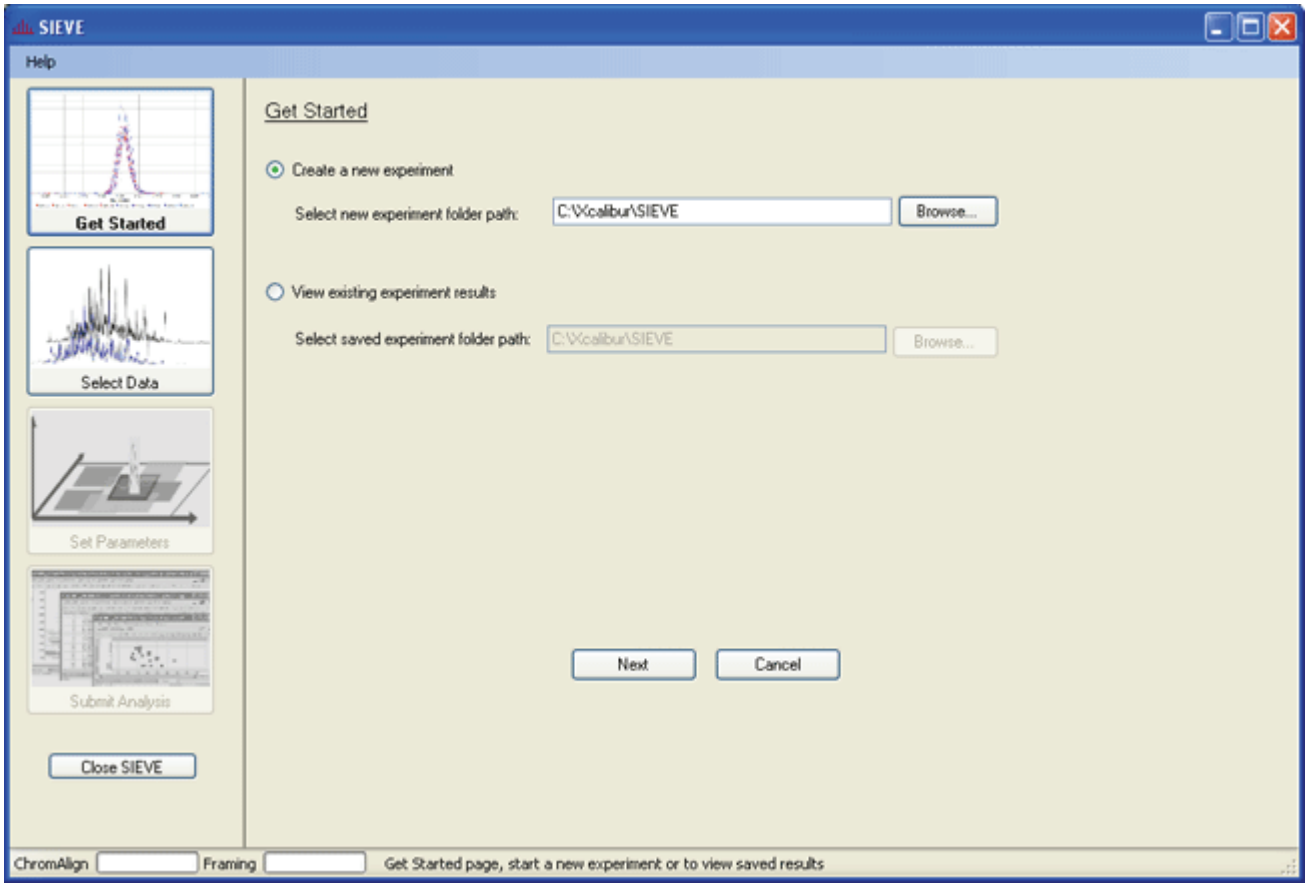
   The following window appears.

   Figure 1. SIEVE home page

2. Click **Get Started** to display experiment options.

   The **Get Started** window appears:

   Figure 2. Get Started window



3. To start an experiment from the beginning, select **Create a new experiment**. To see files from a saved experiment, select **View existing experiment results** (see "Viewing a Saved Experiment" on page 23 for more details about this process).

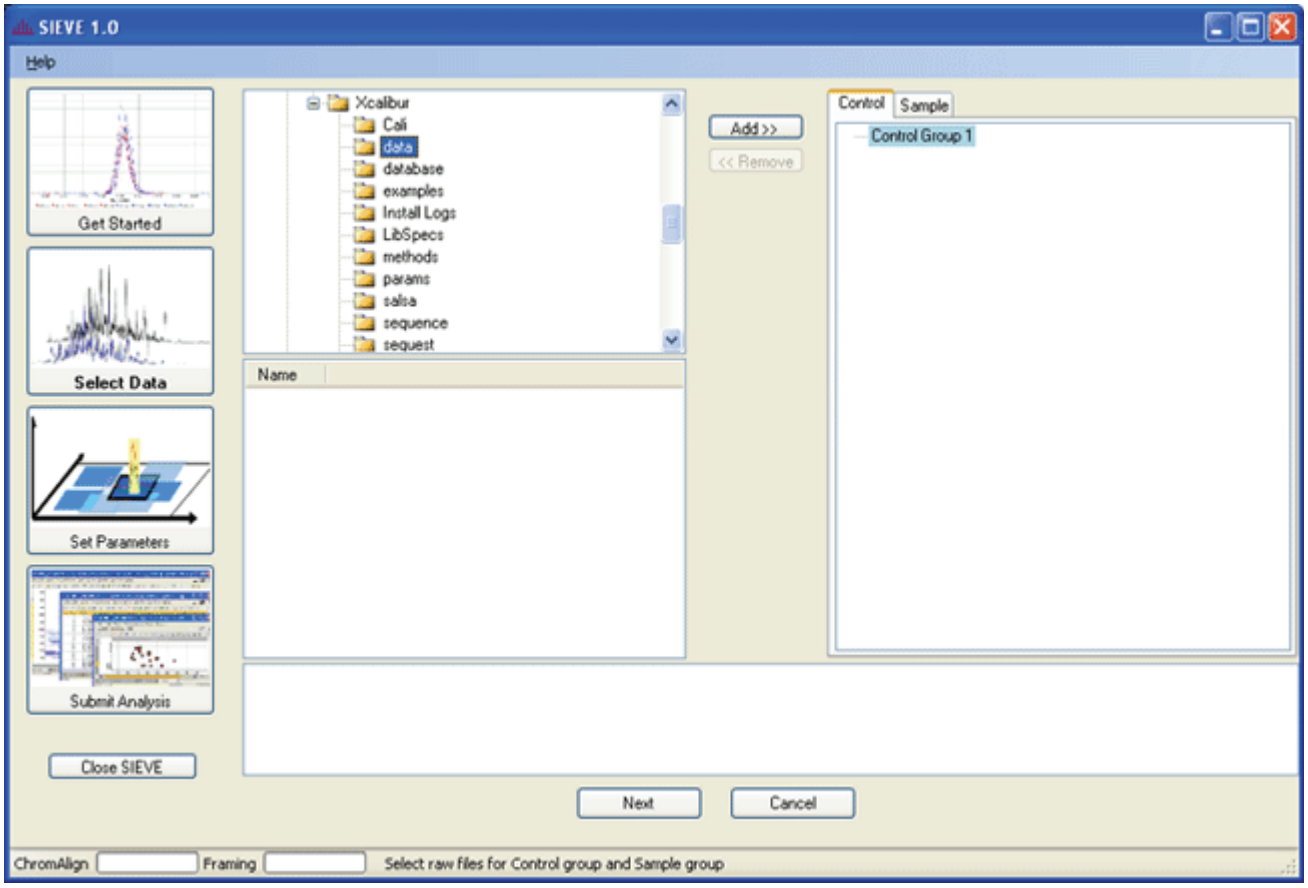
4. Click **Browse** to select the folder path. In the Browse dialog box, click **Make New Folder**.

5. Name the folder and close the browser window.

6. To select control and sample files for the SIEVE process, click **Next** in the Get Started window.

   The following window appears.

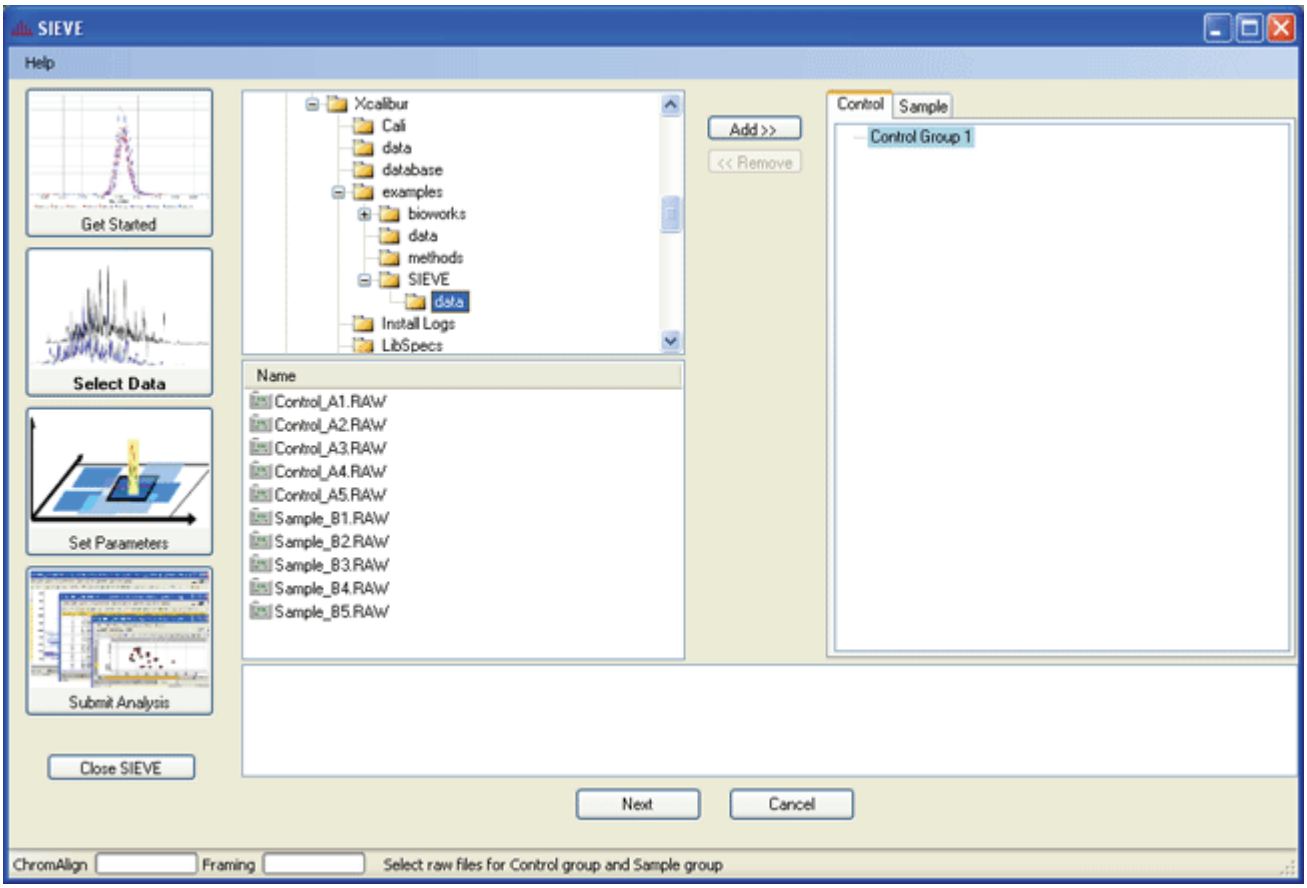
Figure 3. Select Data window



Tip For details about a selected raw file, see the lower portion of the Select Data pane. Review this area to assist in selecting appropriate files for the SIEVE process.

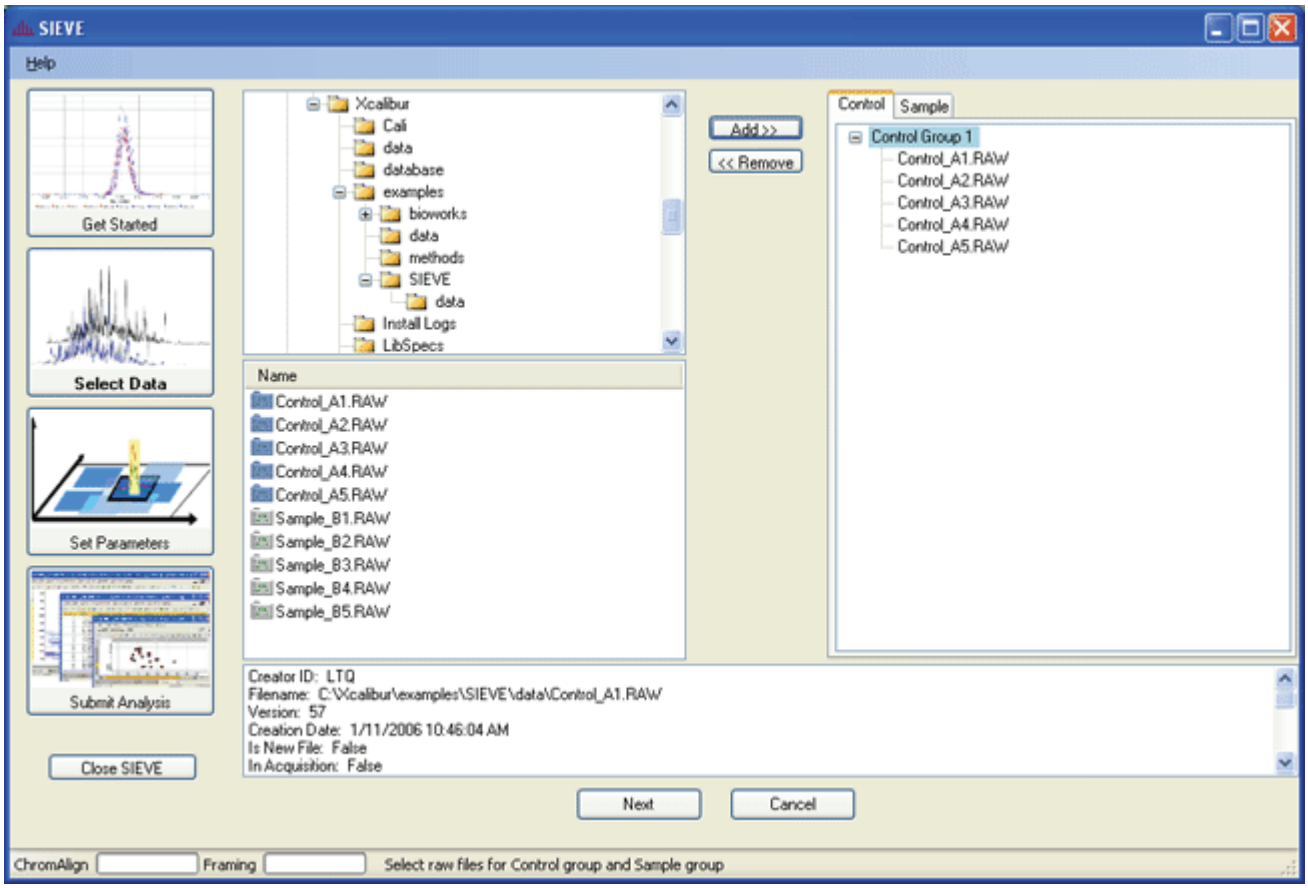
7. Find the Xcalibur folder at the top left side of the window, and click the + sign to open the folders until you see the folder containing your raw files.
8. Click the folder icon (in this case Data) to display the available raw files in the lower left window.

Figure 4. SIEVE folder display



9. To select control files (.raw extension only) from the file list in the lower left side of the screen, click on the file names. Use the CONTROL key to select more than one file at a time.

Figure 5. Selecting raw files from the SIEVE folder display



View the window at the bottom of the screen to find additional raw file information.

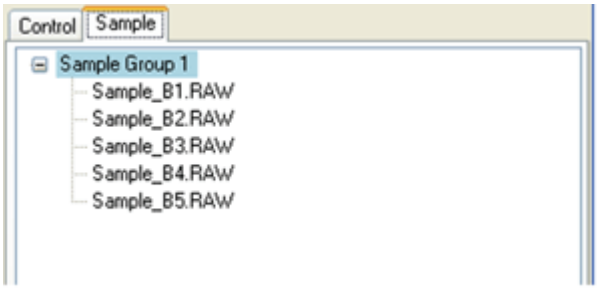
Tip To improve reliability, select control and sample files that have been acquired on the same instrument with the same length of gradient. Do not select the same file for both control and sample.

Tip To achieve optimal alignments, choose a feature-rich chromatographic surface as the first control sample (a chromatogram with good signal-to-noise ratio).

10. Click **Add** to add the files to the Control folder.
11. To add sample files, click the **Sample** tab.

    Select raw files from the file list and click **Add** to add the files to the Sample folder.

Figure 6. Adding sample files



Setting Parameters

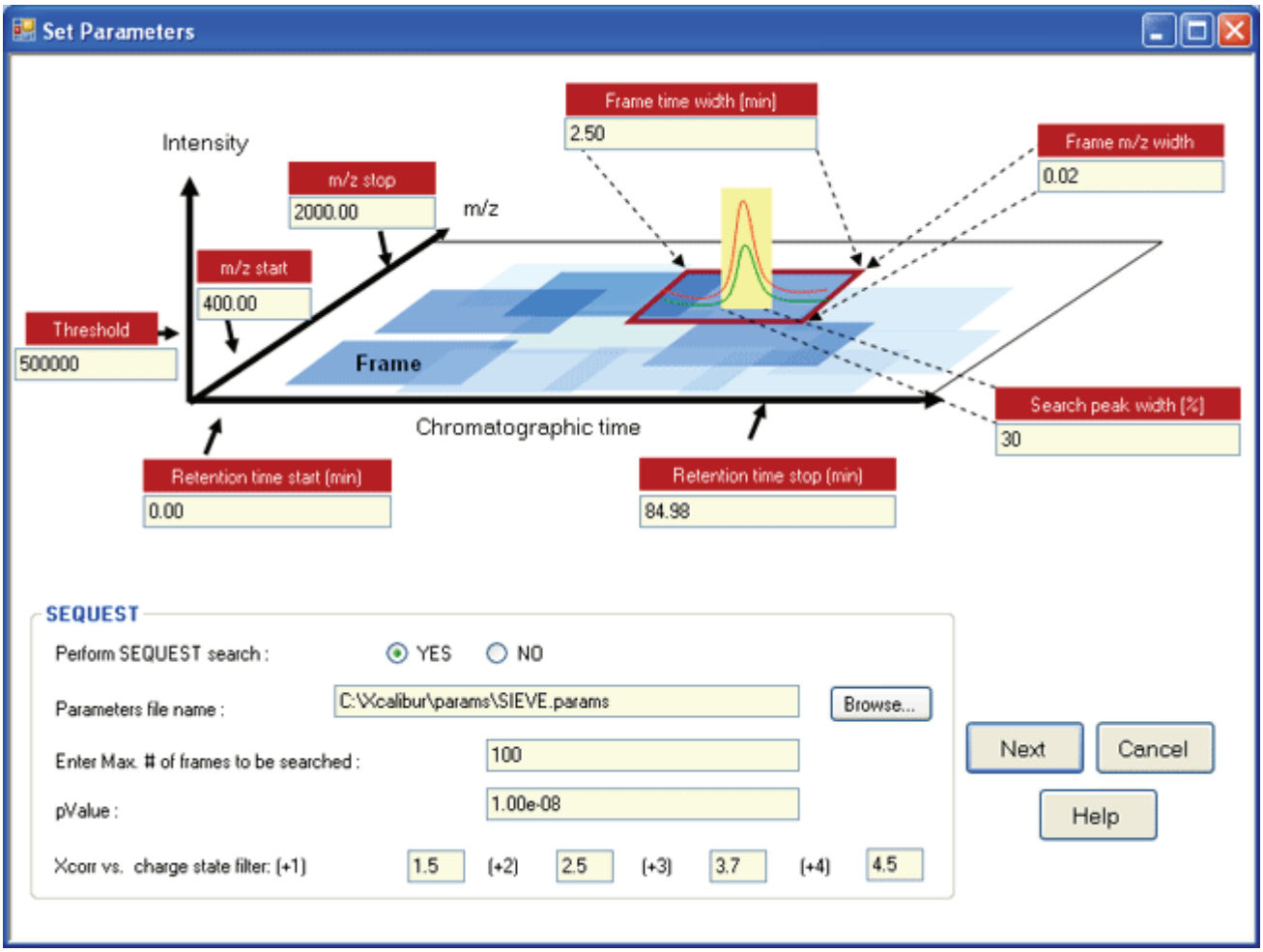
❖ To set parameters

1. After selecting files for the analysis, click **Next** to display the Set Parameters dialog box.

   SIEVE populates each parameter in an edit box with recommended default values based on the instrument used to acquire the first raw file listed in the Control group. You can modify these values from the original default.

2. Change parameters as needed.

Figure 7. Set Parameters dialog box



SIEVE uses these parameters to define a frame to determine the MS intensity threshold and to define the limits of the experiment.

Table 1. Set Parameters dialog box values

Parameter	Definition	LTQ default	LTQ FT and Orbitrap default
m/z start	Type the minimum *m/z* value to be analyzed for each file. Values: +400. The minimum value is the minimum *m/z* of full MS range in the first control raw file. The maximum value is the maximum *m/z* of full MS range in the first control raw file. **IMPORTANT** If you set this value too low, SIEVE might examine the total amount of frames before finding significant peaks	Minimum *m/z* of full range in the first control raw file.	
m/z stop	Type the maximum *m/z* value to be analyzed for each file. Values: The minimum value is the minimum *m/z* of full range in the first control raw file. The maximum value is the maximum *m/z* of full range in the first control raw file.	Maximum *m/z* of full range in the first control raw file.	
Frame *m/z* width	Type a value to define the width of each frame. SIEVE displays a suggested value. Values: 0.01 to 50.0.	1.5	0.02
Frame time width (min)	Type a value to define the width of each frame in minutes. Decrease this number for narrower peaks; increase this number for wider peaks. SIEVE displays a suggested value. Values: 0.1 to 10.0. The default value is 2.5. **IMPORTANT** If you set this value too low, SIEVE might examine the total amount of frames before finding significant peaks	2.5	2.5
Search peak width [%]	This value represents a percentage of the width of each frame starting from the center to define the area where SEQUEST data files (.dtas) are generated for a search. SIEVE display a suggested value. Values: 10 to 100. The default value is 30.	30	30
Retention time start (min)	Type the start of the retention time for the chromatogram in minutes. Values: The minimum value is the minimum retention time in the first control raw file. The maximum value is the maximum of RT in the first control raw file.	Minimum of RT in the first control raw file.	

Table 1. Set Parameters dialog box values, continued

Parameter	Definition	LTQ default	LTQ FT and Orbitrap default
Retention time stop (min)	Type the end of the retention time for the chromatogram in minutes. Values: The minimum value is the minimum retention time in the first control raw file. The maximum value is the maximum of RT in the first control raw file.	Maximum of RT in the first control raw file.	
Threshold	Enter a value to determine the lowest MS intensity that triggers framing. Values: 1 to $1E^{10}$. **IMPORTANT** If you set this value too low, SIEVE might examine the total amount of frames before finding significant peaks	50 000	500 000 for LTQ FT $1E^{6}$ for Orbitrap
Perform SEQUEST search	Click **Yes** to search a protein database using SEQUEST. Values: Yes or No.	Yes	Yes
Parameters file name	Select a SEQUEST parameter file (.params) to be used for a SEQUEST search. See the *BioWorks User's Manual* for further details. The default is C:\Xcalibur\system\programs\SIEVE\params\SIEVE.params.	SIEVE.params	
Enter Max. # of frames to search	Type a maximum number of frames to be searched by SEQUEST. Values: 1 to 10000. The default value is 100.	100	100
pValue	Type the lowest pValue for a frame to be sent for a SEQUEST search. Values: The minimum value is 0. The maximum value is 1. The default minimum value is $1.00E^{-08}$.		
Xcorr vs. charge state filter	Specifies the ID threshold for the filter criteria to improve the number of valid positive entries in the results. This filter combines the peptide properties, using the peptide charge and the Xcorrelation score from SEQUEST. Values: (+1) = 1.5 [0.1 to 10.0] (+2) = 2.5 [0.1 to 10.0] (+3) = 3.7 [0.1 to 10.0] (+4) = 4.5 [0.1 to 10.0]		

Tips for Improving the Process

Use the following guidelines to improve the SIEVE process when comparing files:

- To improve processing speed, narrow the retention time range to the area of the chromatogram where components (for example, peptides and metabolites) have eluted.
- Pick the file with the best chromatogram as the first control file, since all other files are compared to this reference chromatogram.

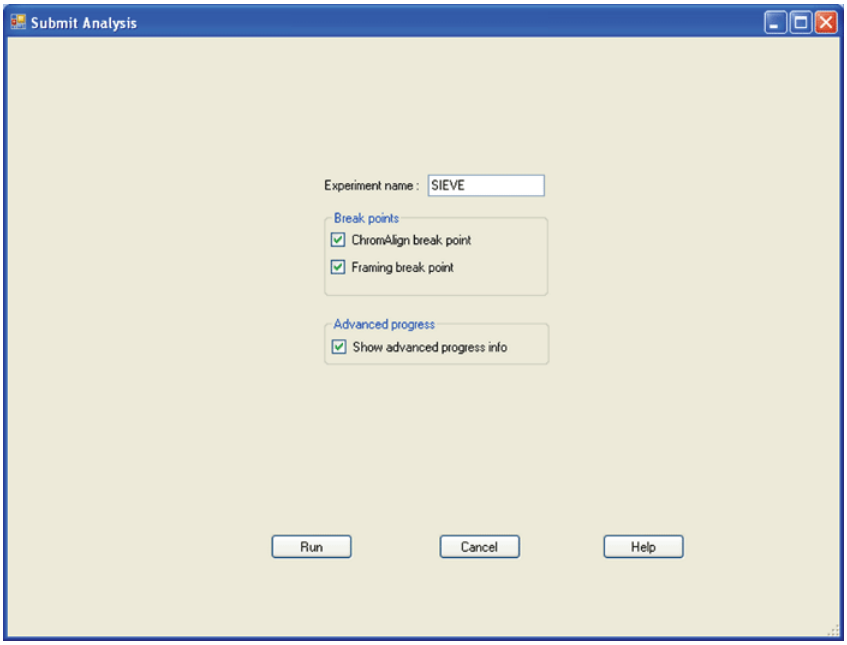
Analyzing Files

❖ To submit files for analysis and display initial results

1. After selecting files and setting parameters, click **Next** on the Set Parameters window to send the raw data from the files to the ChromAlign process. For a description of the alignment process, see Appendix A, "Understanding the ChromAlign Process."

   The Submit Analysis dialog box appears.

   Figure 8. Options for submitting analysis



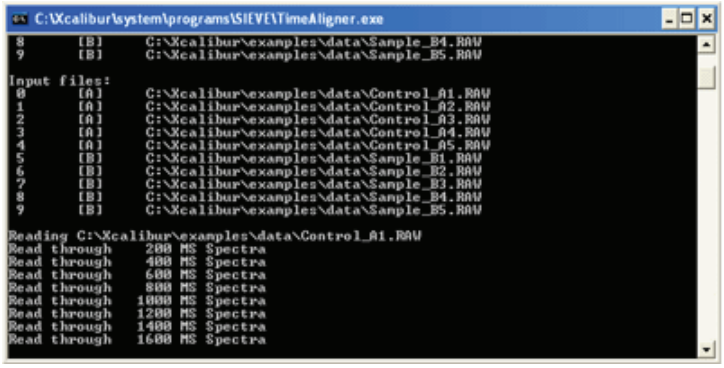
2. Specify a new experiment name and define break points for various steps so that you can evaluate the intermediate results and make changes. See the following table for information about the options.
   SIEVE displays interactive windows for any break points you select. If you don't select any break points, SIEVE finishes the entire process without interruption.

   Table 2. Submit Analysis parameters

Experiment Name	Type the name of the experiment, using the following criteria: • Do not type the name of an existing experiment. • Do not us a name with spaces or special characters. • Use an underscore (_) to separate parts of the file name. If you enter an existing experiment name or an invalid file name, SIEVE displays an error message.
ChromAlign break point	Select this option to pause after the ChromAlign process so that you can evaluate the results. This option opens a Spotfire display after the chromatograms in all control and sample files have been aligned. Use the display to view alignment and to determine whether to continue or to start over with different files.
Framing break point	Select this option to pause the procedure after the framing process so that you can evaluate the results. This option opens a Spotfire report that includes all framing information before you start a SEQUEST search. This report provides all comparative information, including intensity ratios and pValues for each frame.
Show advanced progress info	Click this option to display a DOS window that displays the procedure progress (see Figure 10).

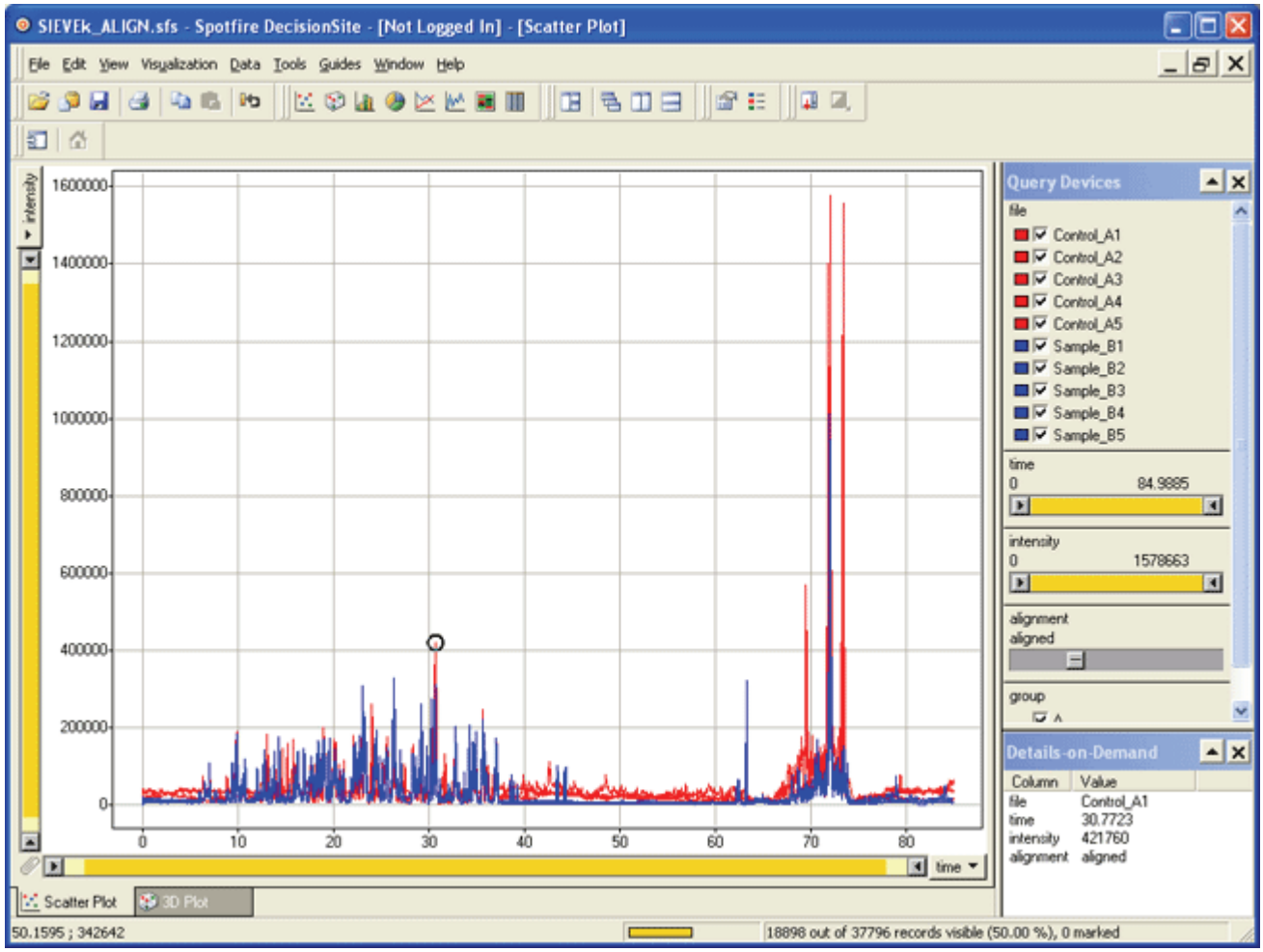
3. Click **Run** to start the analysis and align the chromatograms.

   If you selected **Show advanced progress info** in the Submit Analysis view, a DOS window displays the progress of the procedure.



Spotfire displays a screen similar to Figure 9 with the aligned chromatograms.

Figure 9. ChromAlign results



Framing the Data

Use the following procedure to see the samples in Spotfire, select relevant frames from the Spotfire Gel view, and display them using the SIEVE Framing method. For more information about the framing method, see Appendix B, "Understanding the Framing Process."

❖ To frame the data

1. Close the Spotfire window.

   The Set Parameters dialog box reappears and you can adjust parameters prior to framing. For information about parameter values, see "Setting Parameters" on page 12.

2. Click **Next** to frame the data.

   If you selected **Show advanced progress info** in the Submit Analysis view, a DOS window displays the progress of the procedure.

   IMPORTANT Stay in the SIEVE window when the system is processing. If you click another window or change programs, you might disrupt the processing.

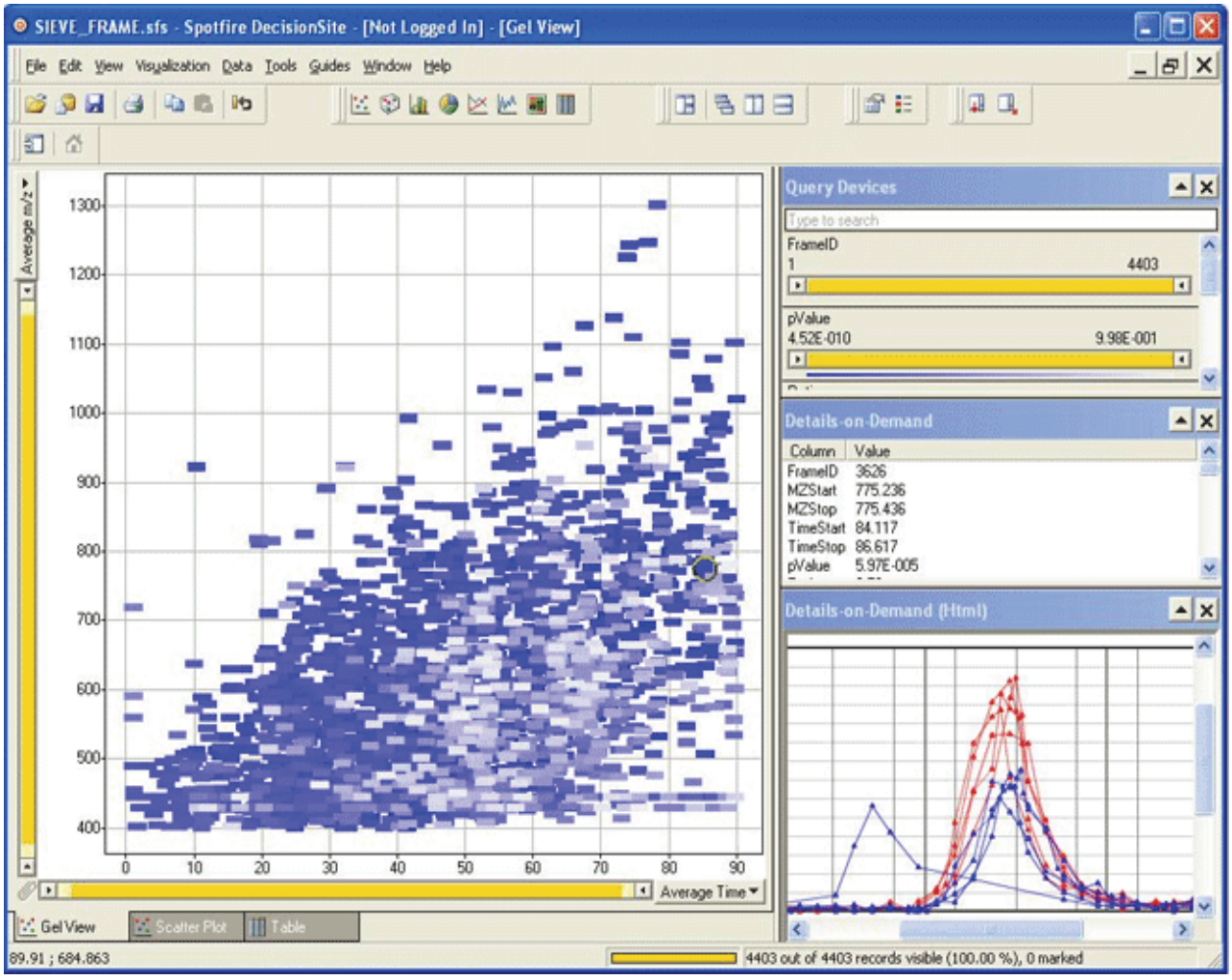
   Spotfire opens, displaying the primary plot.

   Primary plot: Generated by SIEVE within Spotfire, this plot represents frames colored by the pValue to indicate significant differential expression within the frame.

   Consolidated plot: Made up of several hundred to several thousand frames, this plot shows frames of various colors, where darker frames (gels) indicate a more statistically significant difference.

3. To filter down to frames of interest, move the **pValue** slider in the Query Devices section on the right side of the window. If you are looking for a specific ratio, filter by pValue, retention time range, *m/z* range, or even intensity ratios. Low pValues indicate significant changes.

Figure 10. Spotfire view with filtered pValues



For more information about using the Spotfire view, see Appendix C, "Using Spotfire to View Analysis Results."

For further information about seeing and manipulating your data with Spotfire, see Spotfire Help.

Identifying Proteins and Peptides

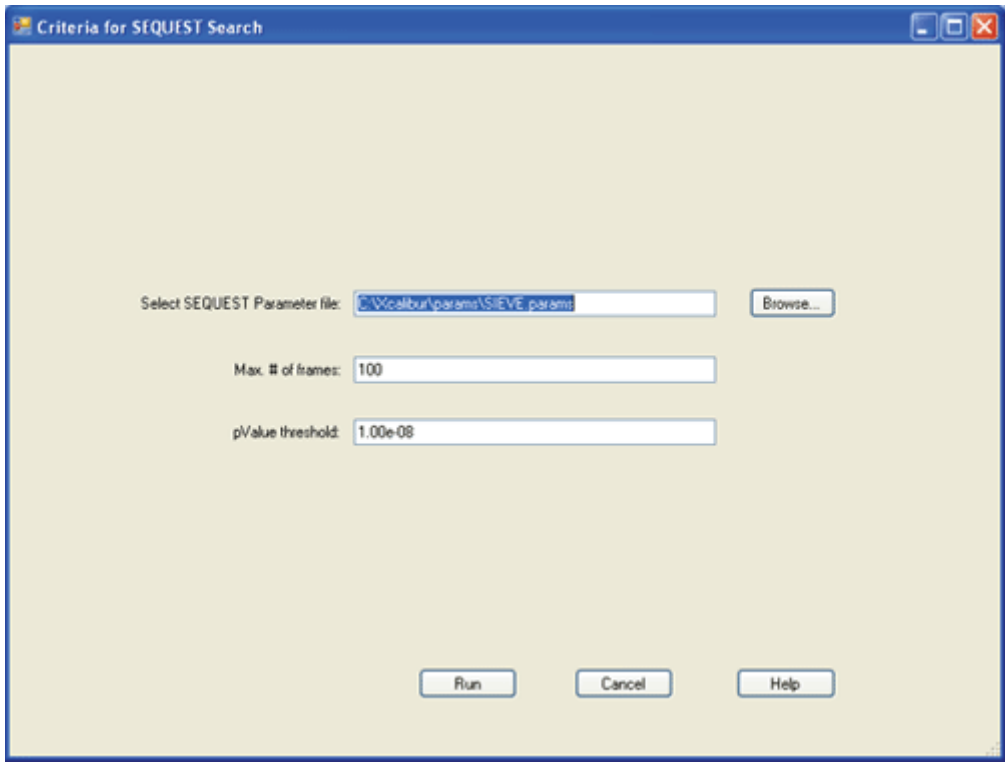
After SIEVE completes the framing process, use the following procedure to send the data to SEQUEST to identify proteins and peptides. After using SEQUEST to perform a database search on the frame, Spotfire displays a table for that frame with information about the protein ID and description, peptide sequence, XCorr, charge state, and peptide identification.

❖ To send the data to SEQUEST

1. Close the Spotfire report.

   The following window appears.

   Figure 11. SEQUEST parameters



   Complete parameter values as follows:

   Table 3. SEQUEST parameters

Select SEQUEST parameter file	To select a SEQUEST parameter file (.params) to be used for the search, click **Browse**. The default is SIEVE.params.
Max. # of frames	Enter the maximum number of frames to search (Values: 1-10 000). The default value is 100.
pValue threshold	Type the lowest pValue for a frame to be sent for a SEQUEST search. Values: 0.0 to 1.0. The default value is $1.00E^{-08}$.

2. Make changes to the parameters as needed. Click **Run**.

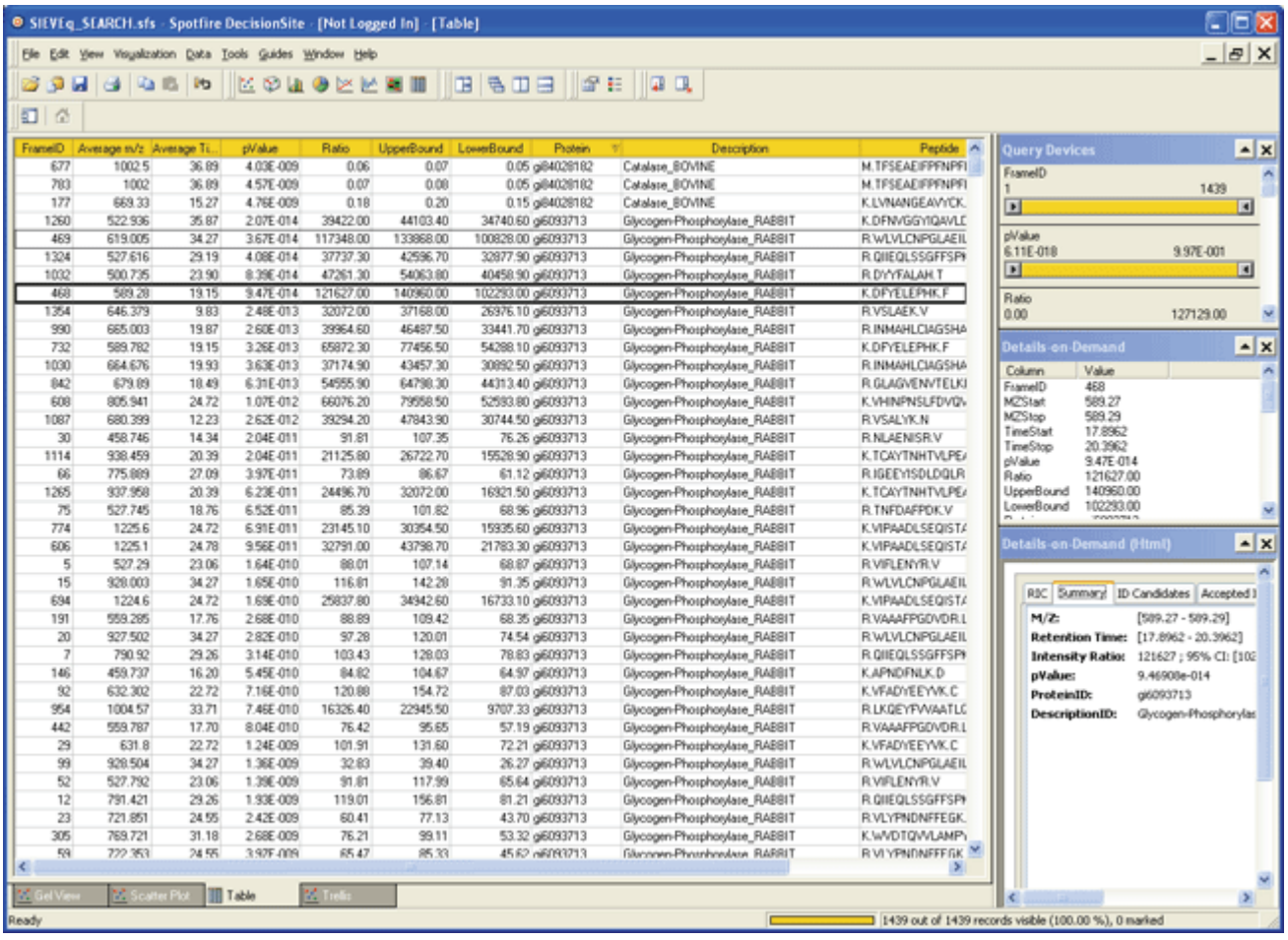
   If you selected **Show advanced progress info** in the Submit Analysis view, a DOS window displays the progress of the procedure.

   At the end of the SEQUEST database search, SIEVE opens the Spotfire application.

3. To see a consolidated summary of SIEVE results, click the **Table View** tab in the Spotfire main view. This table includes columns displaying the pValue and intensity ratio for each frame, as well as the SEQUEST search results, if applicable. SIEVE provides the results of the differential analysis in a table view so that you can check the protein or peptide columns and select a frame with useful protein or peptide information.

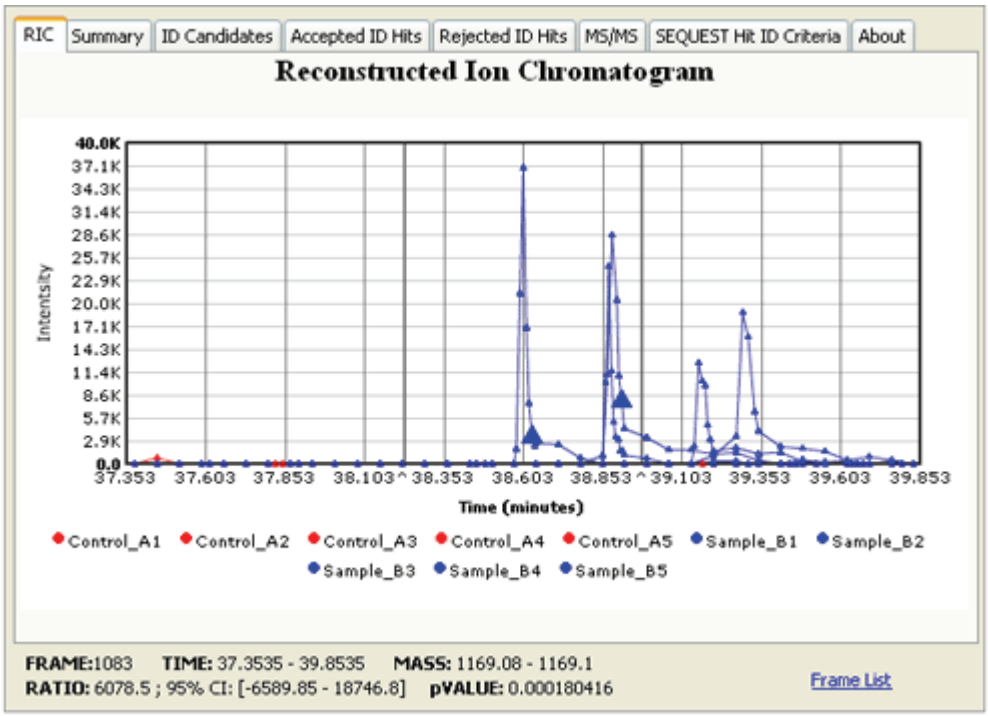
   For information about the various Spotfire views shown as tabs at the bottom of the screen or the sliding scale parameter settings, see Appendix C, "Using Spotfire to View Analysis Results."

Figure 12. Table view in Spotfire post-SEQUEST search



This table is completely interactive. For specific information about a frame, click the entry. Click on a frame in any of the views (for example, a frame in the Gel Spot view or any cell in the Table view) to display this view.

Figure 13. SIEVE detail information display



Note SIEVE does not display raw file names when you select more than 18 control and sample raw files (9 control and 9 sample files).

4. To exit the SIEVE application, click **Close SIEVE** in the SIEVE window.

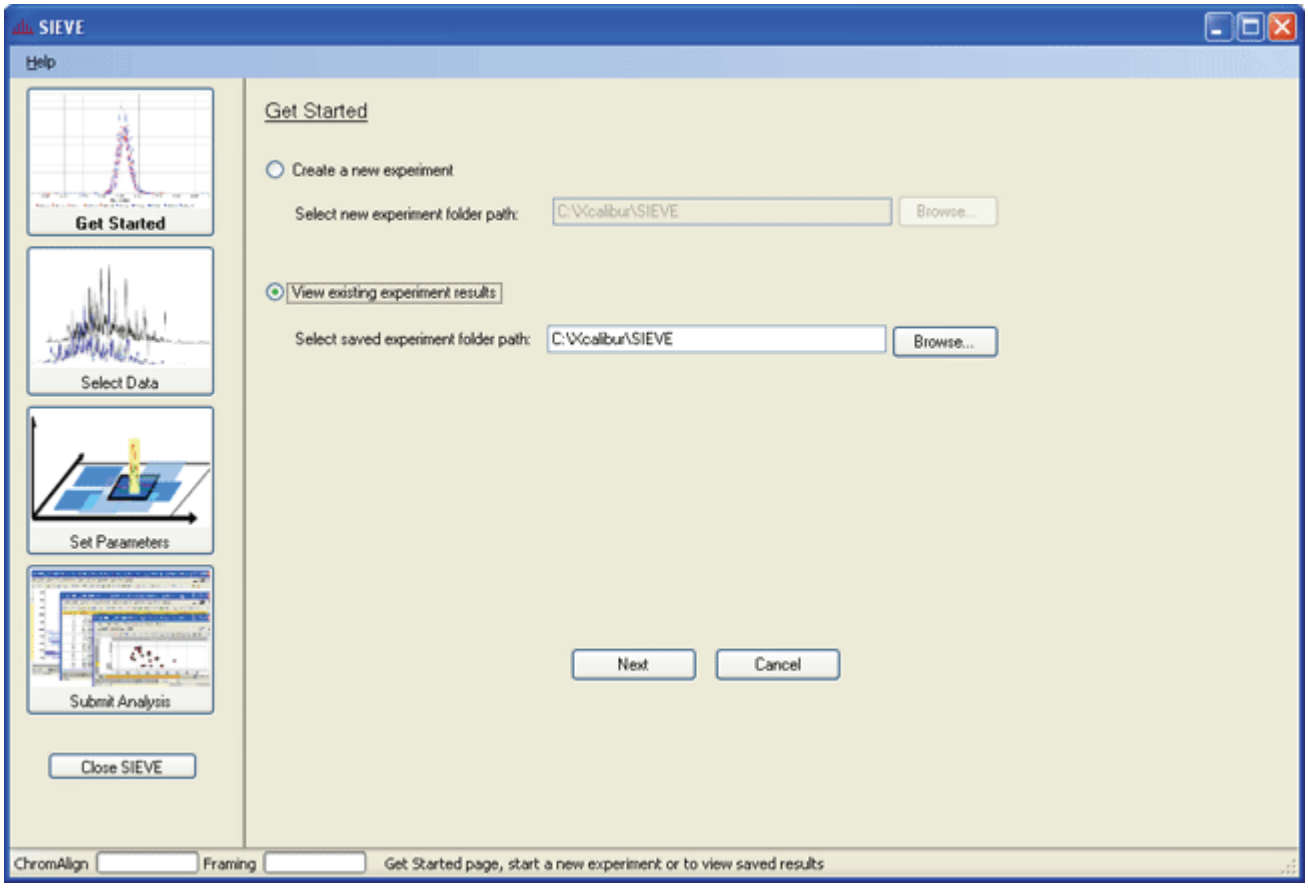
Viewing a Saved Experiment

As you follow the process to create and view a new experiment, SIEVE automatically saves the files that result from the different stages of the process.

❖ To view a saved experiment

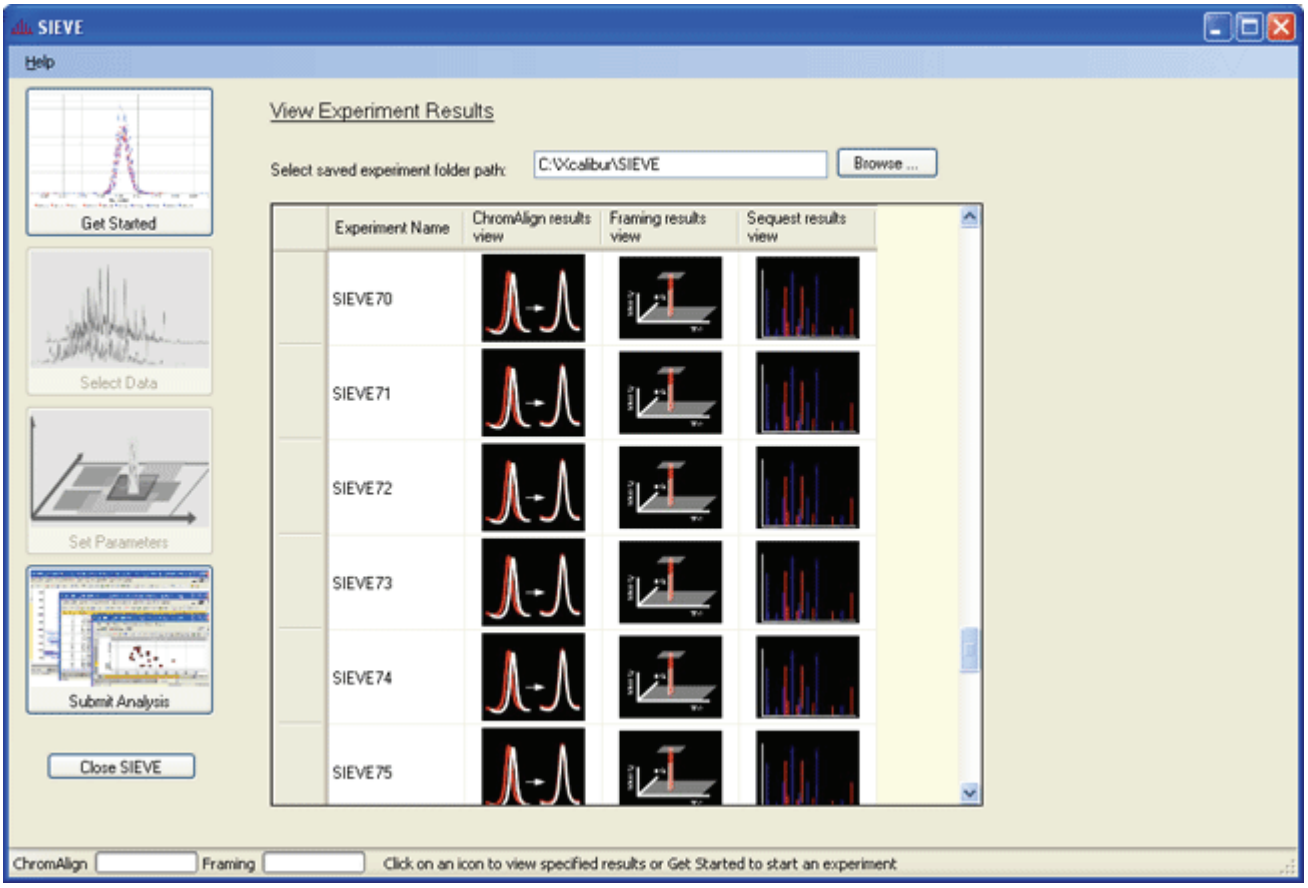
1. To see these results, click **Get Started** and click **View existing experiment results**.

Figure 14. Get Started window



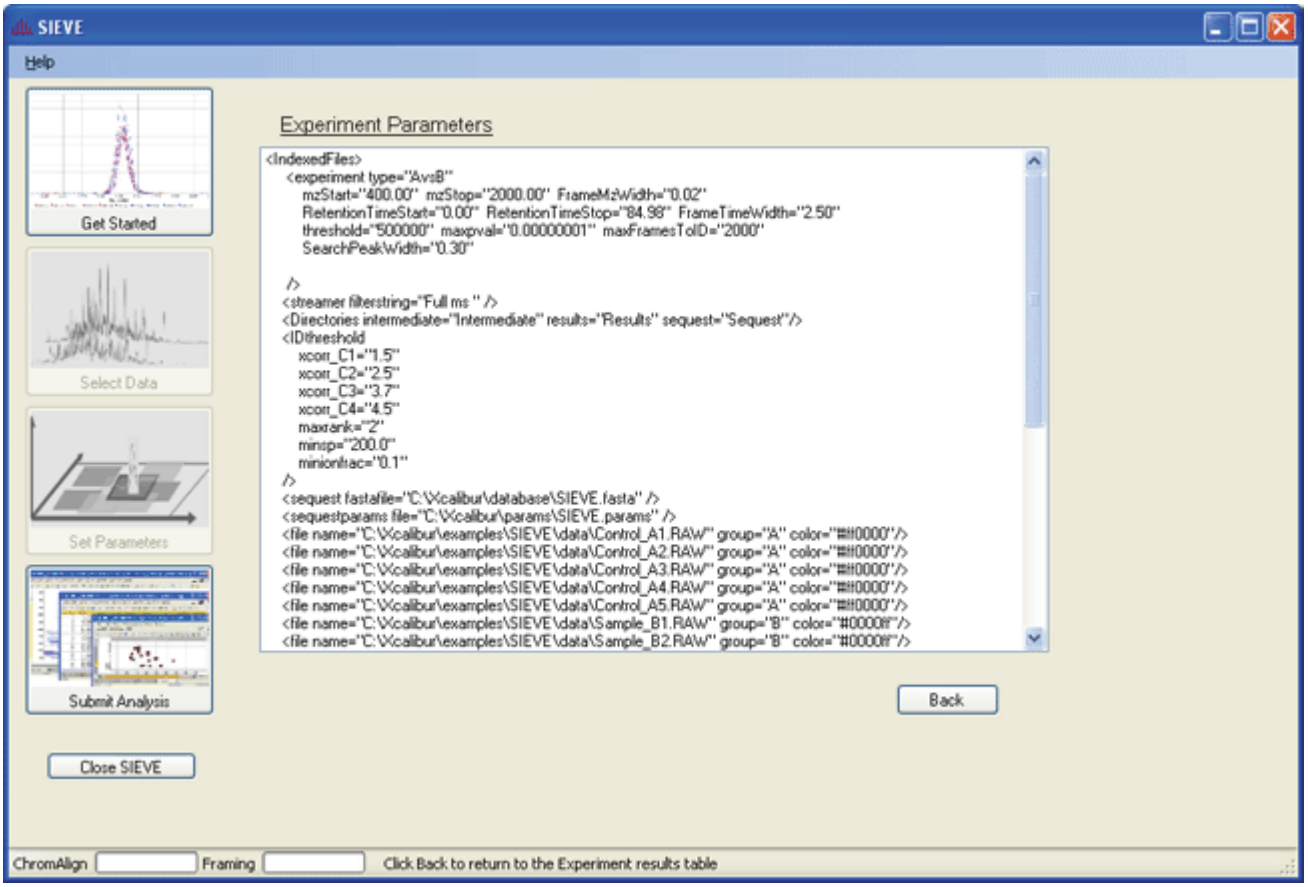
2. To locate the top level folder with experiment results (in this case, C:\Xcalibur\SIEVE\), click **Browse**. Select that folder in the browse window and click **OK** to display the list of possible results. A red x indicates a SIEVE process that was not completed.

Figure 15. View Experiment Results window



3. To see parameters and other details for the experiment, click an icon or the experiment name to see the results.

Figure 16. Experiment Parameters window



4. To return to the View Experiment Results window, click **Back**.
   To exit the SIEVE application, click **Close SIEVE**.

Understanding the ChromAlign Process

When you send an experiment for analysis after selecting sample and control files and setting parameters, SIEVE performs an automatic aligning process called ChromAlign[1]. This process aligns the chromatograms of the sample files with the first control file, correcting chromatographic shifts. The following factors can contribute to chromatographic shifts, which are part of the challenge in biomarker discovery driven by differential analysis:

- The difficulty in reproducing LC/MS experiments
- Aging of separation columns, changes in sample buffer or content, or fluctuations in temperature and sample flow
- The absence of any prediction model that explains peptide elution in chromatographic columns and allows alignment of eluting peaks through the accurate modeling of peptide elution.

The following sections provide detailed information about the ChromAlign process:

[1]ChromAlign: US Patent applied for.

Outlining the ChromAlign Process

Start with Original Chromatograms A and B

After selecting samples and determining the control data set, send the control and sample chromatograms to be aligned with respect to the first control chromatogram.

Step 1
Align chromatograms roughly using Fast Fourier Transform (FFT) algorithm

Step 1 of the ChromAlign process produces a crude alignment of the two chromatographic surfaces using a fast fourier transform algorithm to maximize the overlap of the chromatographic surfaces:
FFT (A) (first control sample)
FFT (B) (all other chromatograms)
FFT^{-1} (FFT (B)xFFT (A)*)
Crudely aligned, B^-

Step 2
Generate correlation matrix, (C (A, B))

In step 2, SIEVE uses the full MS scan information to fully align all components (major as well as minor components of the samples) of the chromatographic surfaces. SIEVE uses dynamic programming to maximize the sum of all correlation co-efficients of peaks between time 0 and the end of the chromatogram.

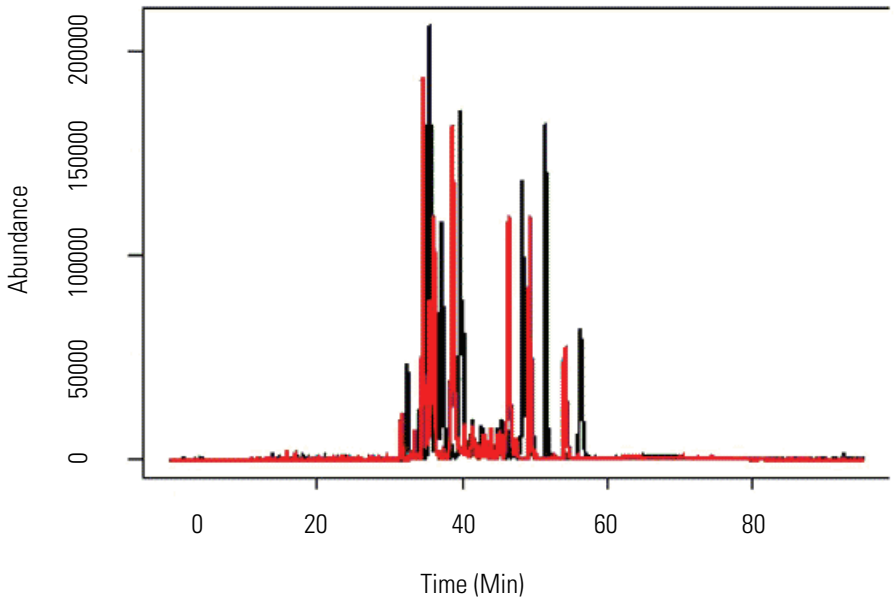
Step 3
Create dynamic programming optimal path in $C(A, B^-)$ and provide a final alignment

SIEVE completely aligns the chromatographic surfaces.

Starting the ChromAlign Process

After you submit the data for analysis, SIEVE aligns the chromatographic profiles (using the base peak information only) using fast fourier transforms (FFTs) to provide a transformation-based correlation analysis between the two chromatograms (see Step 1). This alignment produces a correlation array showing exactly how much to shift the various chromatograms to align the major features. Figure 17 shows original chromatograms before the ChromAlign process begins. In Figure 17, the black trace is the trace of the control sample and the lighter (red) trace is the test sample. The ChromAlign process aligns the test sample to the control sample.

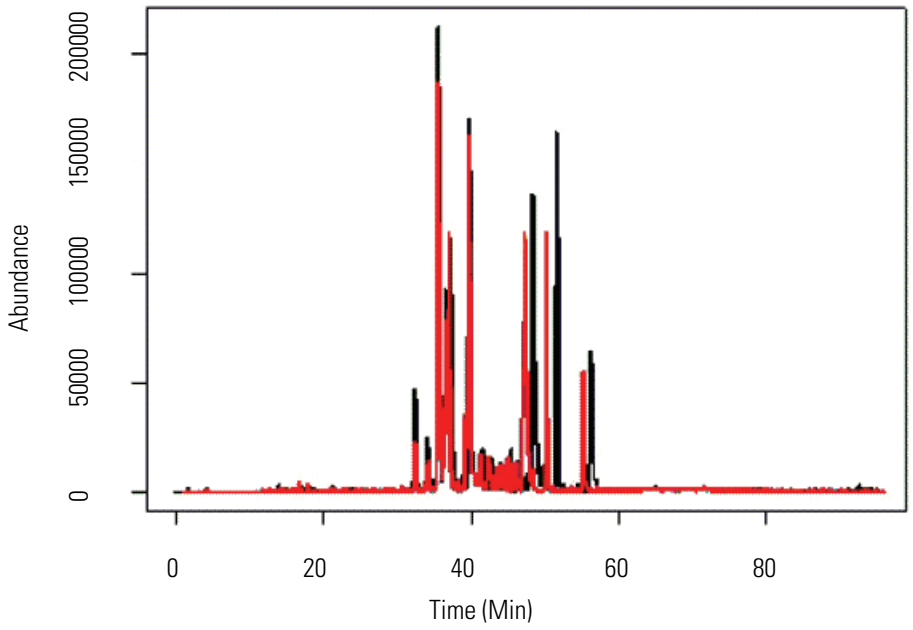
Figure 17. Original chromatograms



ChromAlign First Level Alignment

Figure 18 shows sample results from the first level ChromAlign process. Some of the major peaks in the red trace are now aligned with respect to the control sample (for example, see the peaks at around 35 minutes). SIEVE supports comparing 50 chromatograms to another 50 chromatograms, totaling 100 raw data files. SIEVE aligns all chromatograms to the first control file that you chose when originally selecting control and sample files.

Figure 18. ChromAlign first level processing



The following problems can complicate the alignment:

- Both linear and non-linear shifts in the time axis
- Variations in the intensity
- Dynamic-range challenge when aligning low abundance ions in the presence of high abundance ions

Note Choose a feature-rich chromatographic surface (a chromatogram with good signal-to-noise ratio) to achieve optimal alignments.

Figure 18 shows the alignment of two chromatographic runs of a digested nine-protein mixture: horse myoglobin, bovine serum albumin, chicken egg lysozyme, chicken egg ovalbumin, bovine carbonic anhydrase, bovine ß-casein, horse cytochrome C, bovine α-lactalbumin, and rabbit glyceraldehyde-3-phosphate dehydrogenase.

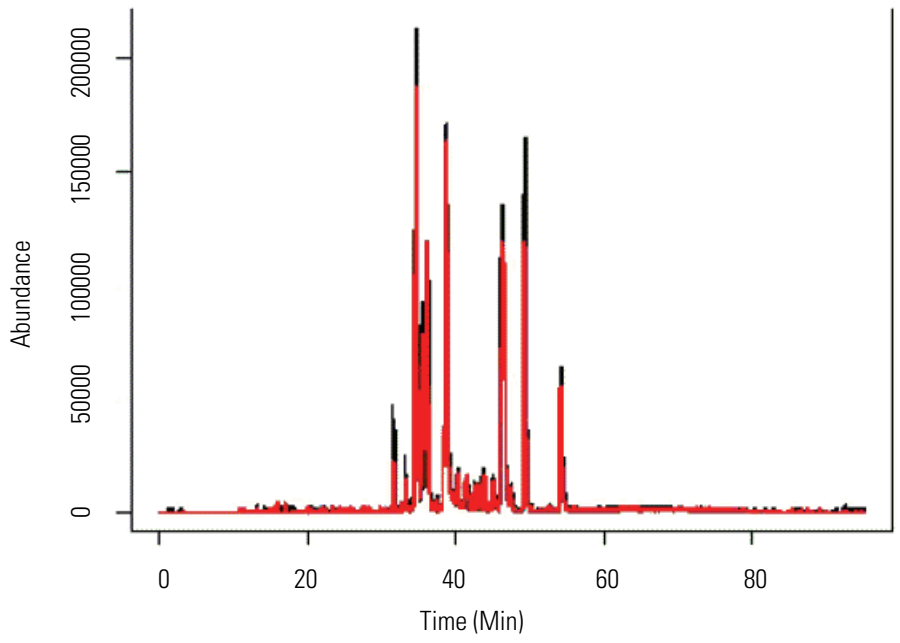
ChromAlign Second Level Alignment

The second step of the ChromAlign process refines the alignment of the roughly aligned chromatograms. ChromAlign generates a correlation matrix between full MS scans of the two chromatograms, populated by the Pearson co-efficients obtained from the full MS data. Then it assigns artificially large values to the correlations between the first scans and the last scans, respectively (see Step 2). ChromAlign then aligns the first and last scans, connecting these points to maximize the sum of the correlation co-efficients (see Step 3).

Dynamic programming generates the optimal path (maximum sum of the correlation co-efficients) to yield a chromatographic time warping function (see Aligning Over Chromalign Time Shifts). Applying this function to the chromatographic surface of the sample produces full alignment of both data sets (see "Aligning Over Different Abundance Levels" on page 32).

After final alignment, SIEVE displays the optimal path in the form of the aligned chromatographic times and provides the alignment score (the average of the correlation matrix along the generated optimal path). Figure 19 shows the aligned results for the same chromatograms where all components of the nine-protein mixture are fully aligned.

Figure 19. ChromAlign second level processing

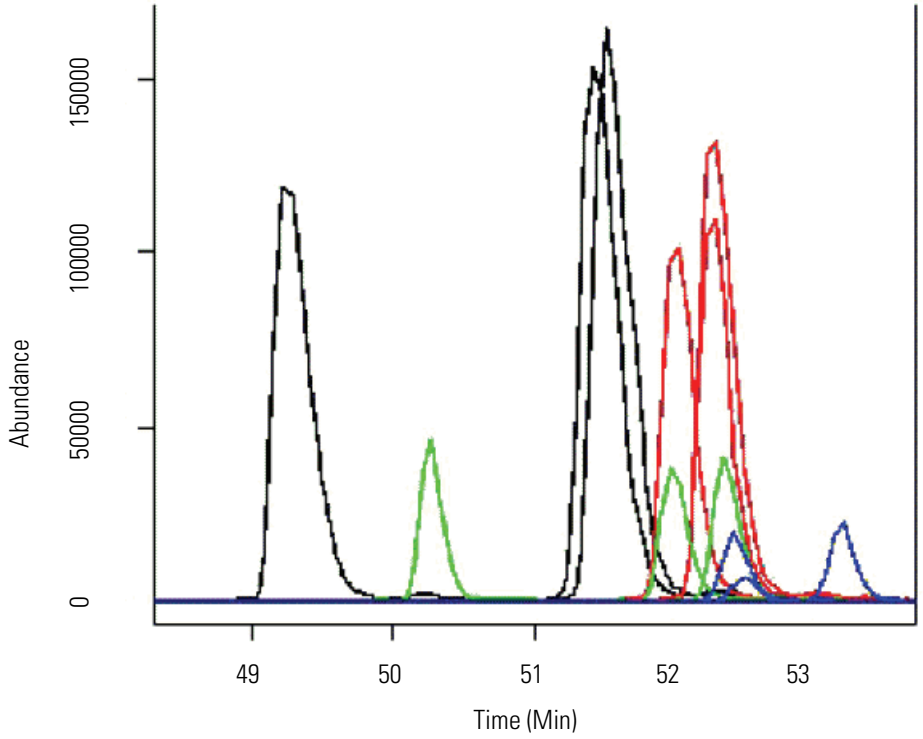


This example demonstrates good alignment after the ChromAlign process, even though the alignment is not a simple, static shift.

Aligning Over Different Abundance Levels

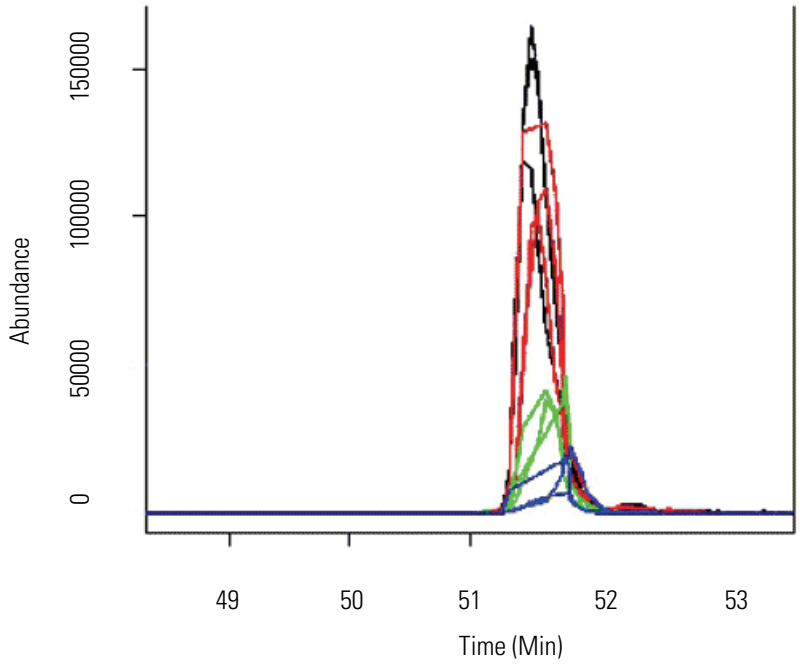
Biomarker discovery detects low abundance components in the presence of high abundance components that remain essentially unchanged between the controls and samples. The following example shows the alignment of low abundance components in the presence of high abundance components. Figure 20 shows reconstructed ion chromatograms of a horse myoglobin peptide added to the nine-protein mix at three different concentrations (500 fmol [red], 250 fmol [green], and 100 fmol [blue]) and compared to the control sample (1000 fmol [black]).

Figure 20. Original reconstructed ion chromatograms



ChromAlign produces perfect alignment of the 100 fmol trace with the 1000 fmol trace.

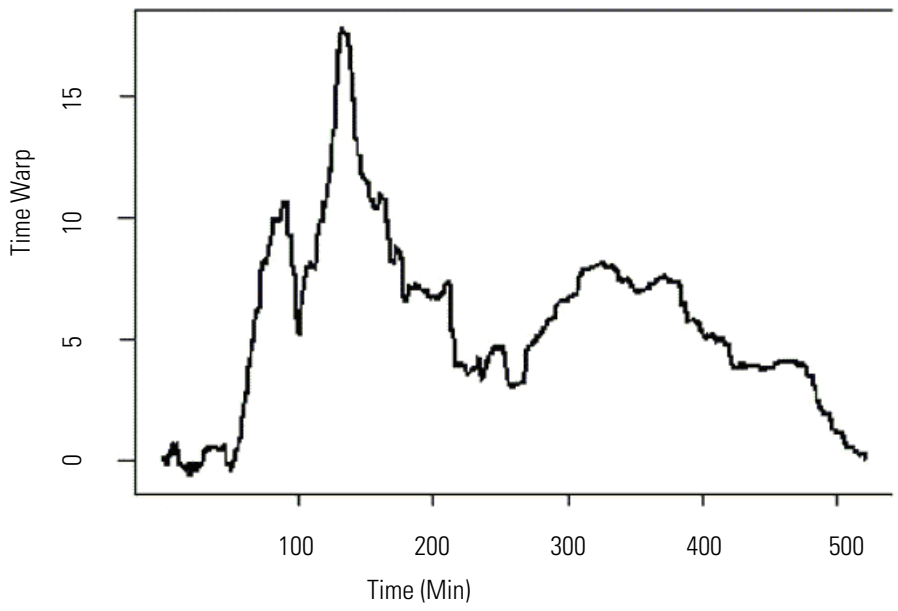
Figure 21. Chromatograms after alignment



Aligning Over Chromalign Time Shifts

See Figure 22 to understand the complexity of the alignment process by plotting the chromatographic time warping function. Although this plot uses data from a different set of control and sample files, it demonstrates the complexity of the alignment effort caused by chromatographic shifts. The average alignment score for the alignment in this figure is 0.53.

Figure 22. Chromatographic time warping required to create alignment



Understanding the Framing Process

To understand the framing process in SIEVE, consider that biomarker or differential expression experiments require comparing at least two sets of raw data files, each representing two distinct biological states (A compared to B); however, to be statistically meaningful, a well-designed biomarkers study should aim to compare multiple bioreplicates from control and treatment (disease) sets, each analyzed, ideally, multiple times as technical replicates. Furthermore, presenting the results of the typically large volume of generated data in one consolidated view improves the differential analysis.

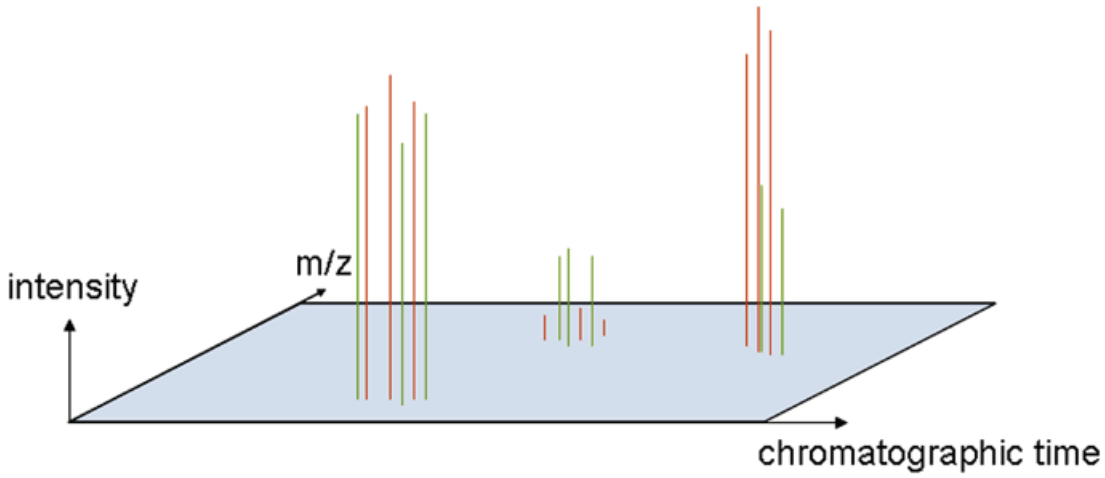
This discussion of the framing process covers the following topics:

- Aligning Peaks
- Finding the Highest Peaks

Aligning Peaks

In this simplified representation of an experiment comparing three raw files (Controls) to three raw files (Treatments), SIEVE first consolidates the spectra in a multivariate space: Retention time versus *m/z* versus full-scan ms intensity (see Figure 23). Each of the control samples is represented by green spectral peaks on the map; each of the treatment samples is represented by red spectral peaks on the map. (In a real set of samples, thousands of peaks would populate this plane).

Figure 23. SIEVE consolidates the spectra

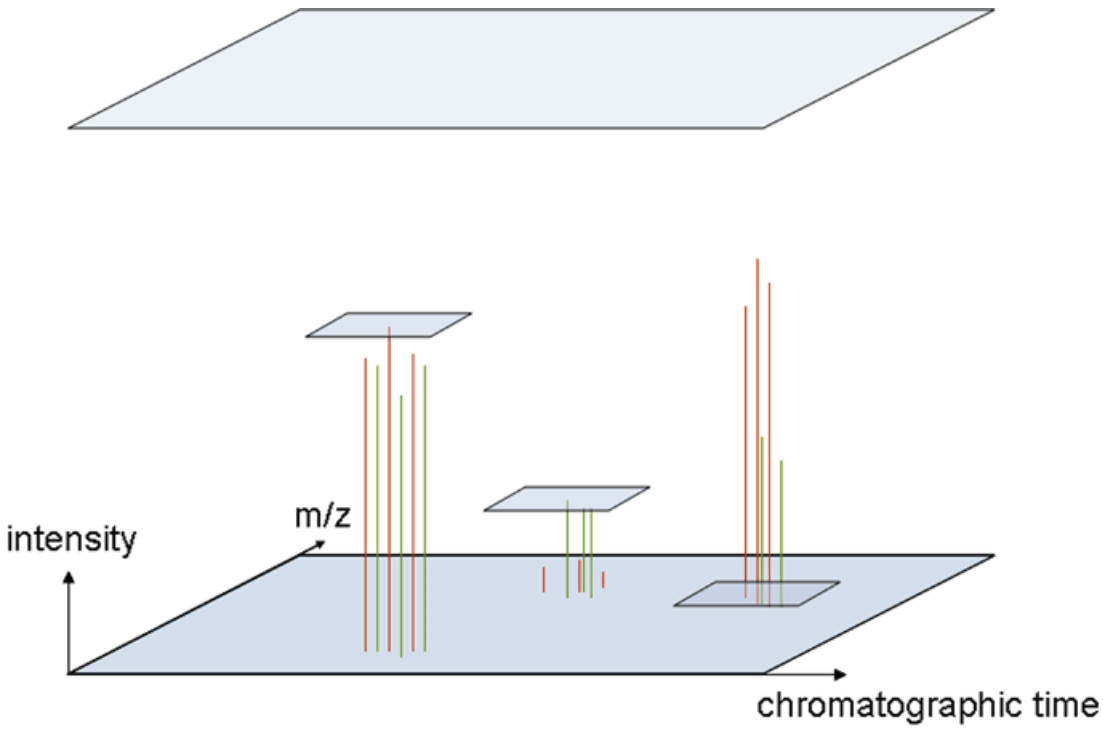


Notice that there are green outlying peaks to the left of each of the three main clusters. These peaks represent chromatographic shifts in retention time, requiring that the peaks first be aligned using the ChromAlign process.

Finding the Highest Peaks

After alignment is complete, framing begins with the most intense MS peak. To find the first most intense peak, imagine a flat plane being lowered vertically over the 3D plot until a cluster of peaks with the highest intensity is found, triggering the creation of a frame. This iterative process is similar to the way data-dependent scanning operates in which the intensities of the MS peaks trigger MS/MS events. Using a proprietary algorithm (Recursive Base Peak Framing), SIEVE draws a frame around this cluster, which includes peaks from both the control and treatment samples (see Figure 24). The width and length of each frame depends on user-defined parameters selected before starting the SIEVE process, but the program suggests default values based on the type of instrument used to acquire the data.

Figure 24. Finding peaks



Using Spotfire to View Analysis Results

The following sections describe how to use Spotfire to evaluate your experiment results and improve the process:

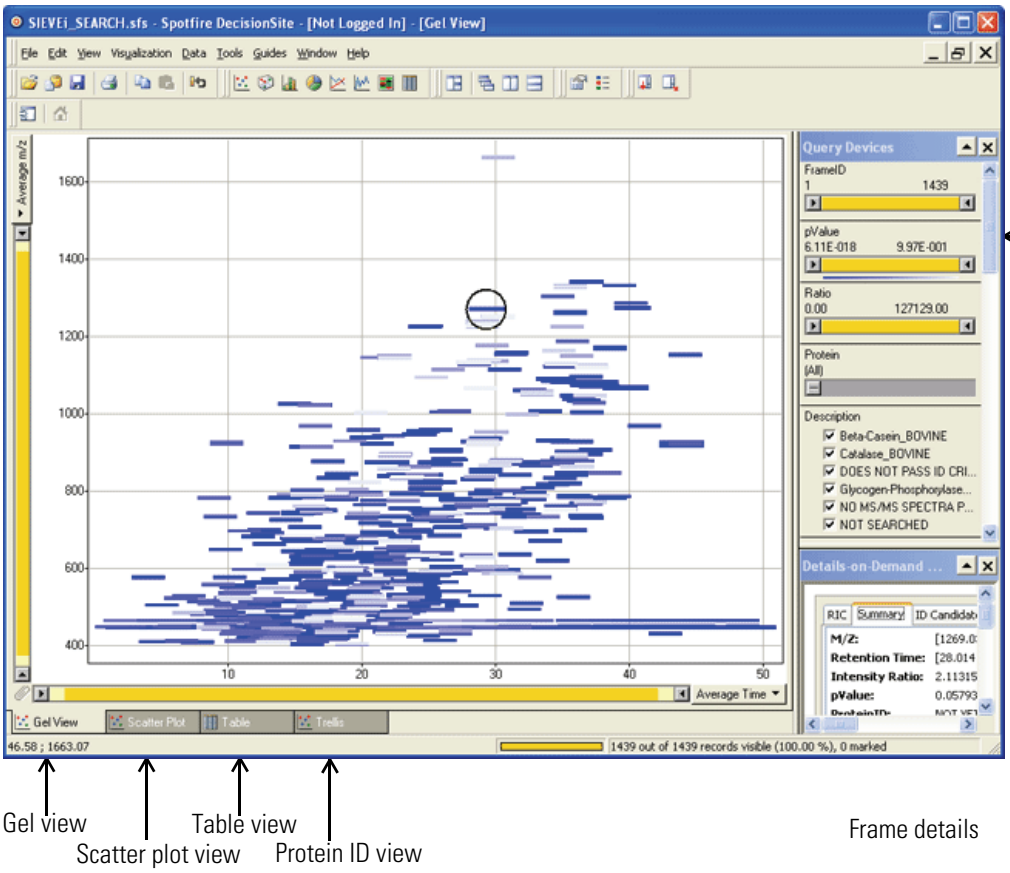
- Using the Spotfire View
- Spotfire Views

Using the Spotfire View

You can use the Spotfire view to review and fine tune the data. The Spotfire view is interactive, but if you make changes in the Spotfire view, it does not make changes to your original experiment parameters.

Figure 25 shows the basic elements of the Spotfire main window.

Figure 25. Initial Spotfire view



You can keep several Spotfire windows open at the same time (up to three), but if you click **Get Started**, SIEVE closes all the Spotfire windows. You can do any of the following:

- Adjust query values to exclude frames that do not match search criteria. For example, to adjust the slide range for pValue to include pValues that are closer to 0, slide the rightmost arrow away from 1.0.
- To display frame information on the Spotfire page, select a frame by clicking on a gel in the Spotfire Gel view.
- View the data in several views (see "Spotfire Views" on page 42).

Spotfire Range Sliders

Use the Range sliders to select a range of values, to filter the data, or to select data subsets of interest. Use the left and right arrows to change the lower and upper limit of the range, displaying and selecting only records with values within the chosen range. Labels above the slider indicate the selected span. The slider automatically jumps to values in the data set (not necessarily the visible or selected records).

Click and drag the yellow portion of the range slider to pan the selected range. Observing the response of the other sliders to such a sweep can give some interesting clues about the correlation between parameters in the data set. For more information, see Spotfire Help.

Frame Details

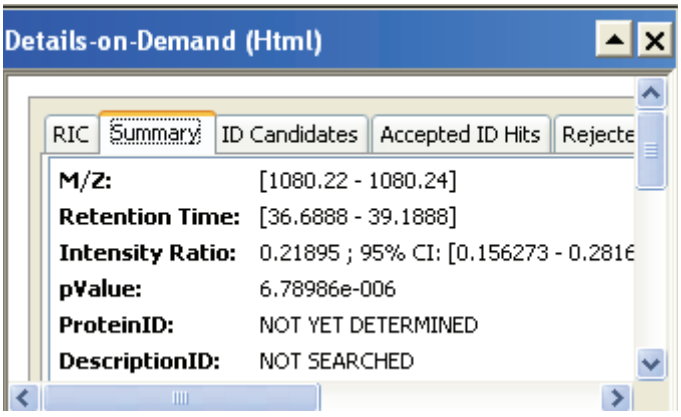
The **View Details on Demand** (text) option from the Spotfire View menu (default) displays data in the lower right side of the Spotfire display. After selecting a frame from the gel view, use this data to decide whether the frame is useful for further examination.

Gel Cell

Click a gel in the Spotfire window to select it (a circle surrounds it) and display the following:

- Reconstructed ion chromatograms for that frame
- The mass-to-charge ratio around which the frame was centered, the ratio of peak intensities between the control and treatment samples
- The confidence interval
- The pValue, indicating whether the ratio was statistically significant

Figure 26. Spotfire detail for a frame specified by pValue



Spotfire Views

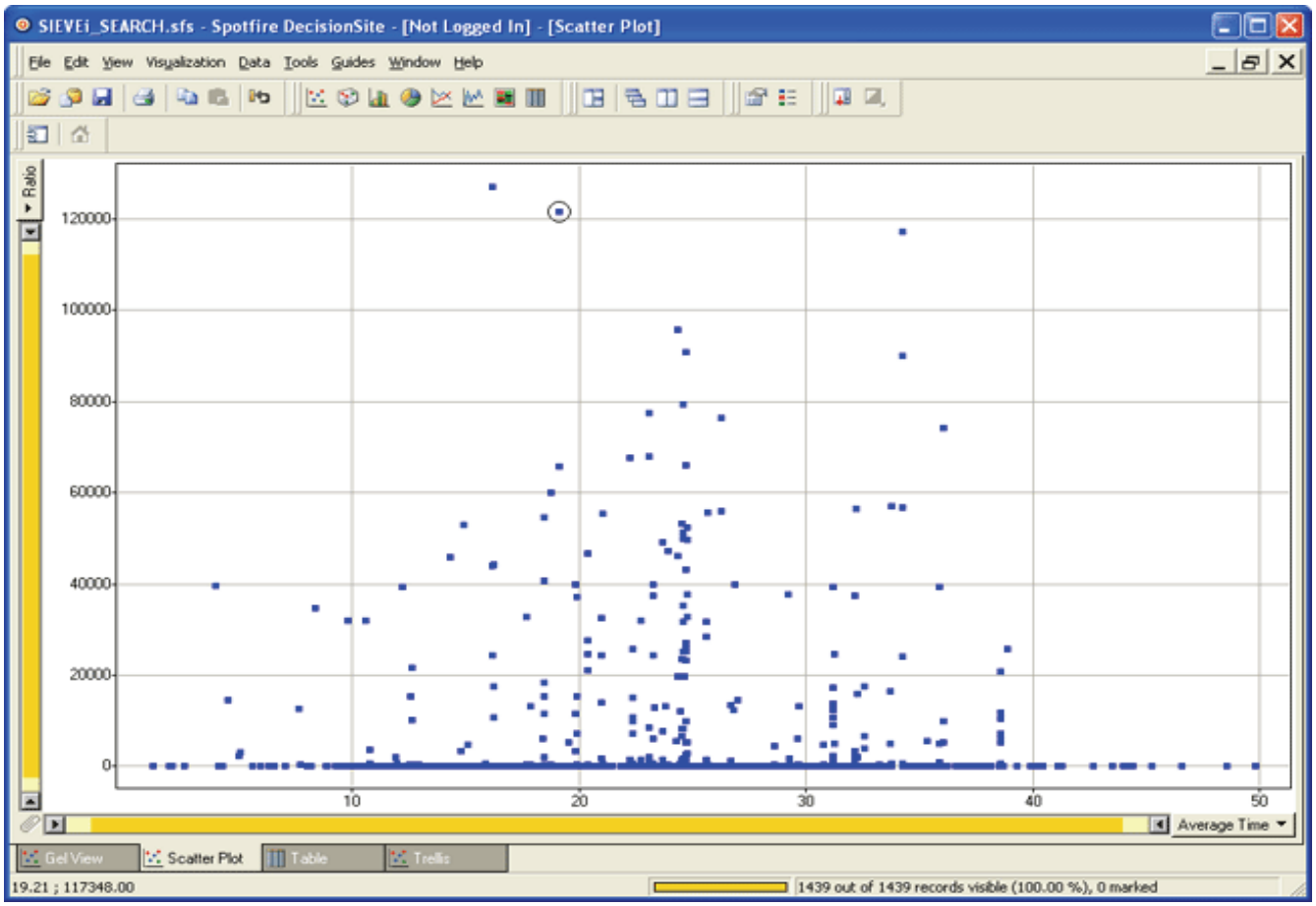
Other Spotfire views include the following:

- Scatter Plot View
- Protein ID View
- Results Summary Table

Scatter Plot View

The scatter plot view is the quickest way to review the significant differences in a biomarker discovery experiment. This view in SIEVE displays all of the peptide expression ratios on a single plot of ratio versus retention time.

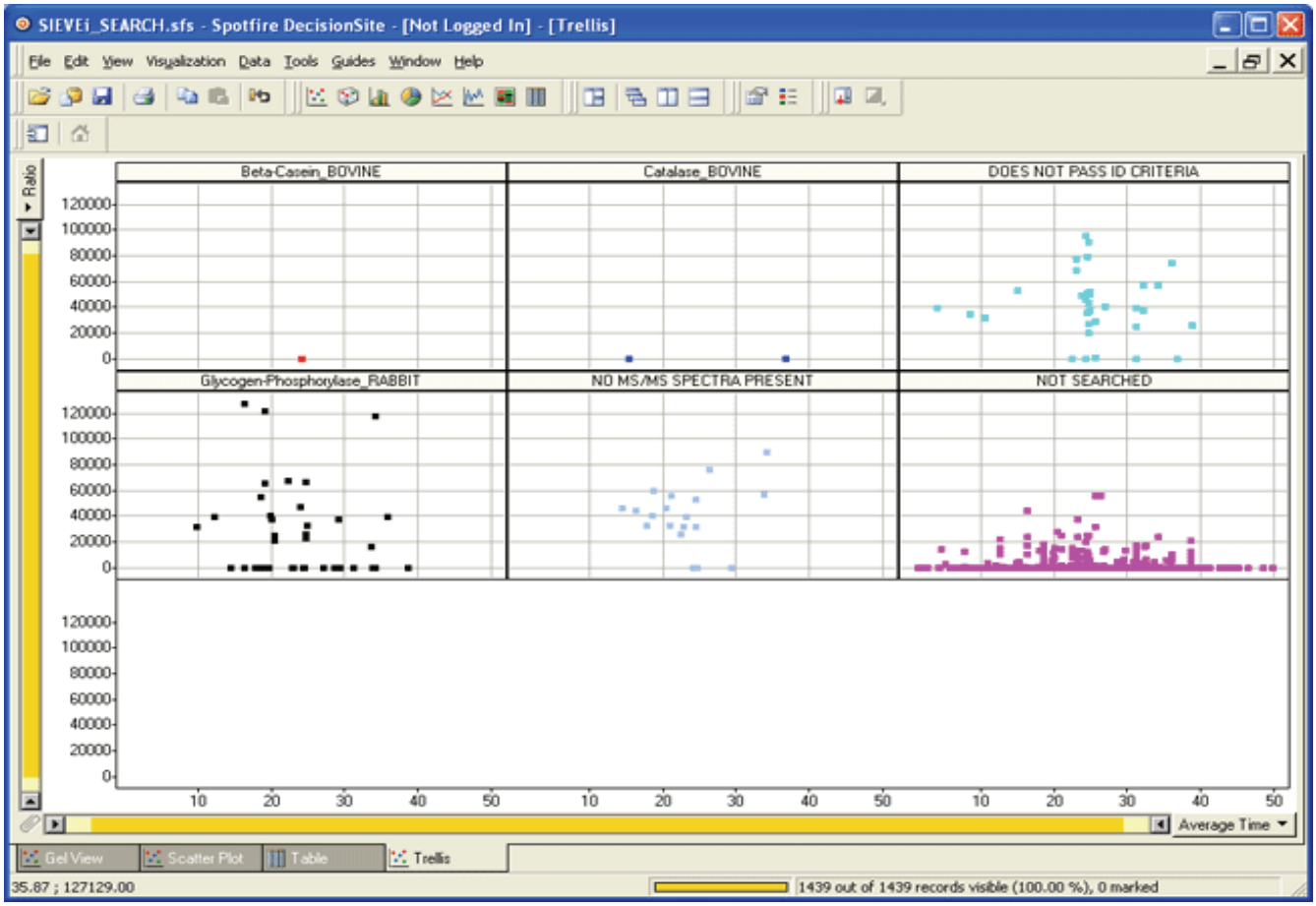
Figure 27. Spotfire Scatter Plot view



Protein ID View

Each protein can also be represented by an individual scatter plot, providing a means to quickly scroll through the expression ratios for each protein identified from A to Z.

Figure 28. Spotfire Protein ID view



Results Summary Table

SIEVE presents the results from an entire biomarker discovery experiment in an easy-to-interpret interactive table. Sort data by ratio of change, pValue, protein identification, *m/z* value, or XCorr, for example, with the additional flexibility of modifying columns to suit individual experiments.

Figure 29. Spotfire Results Summary table

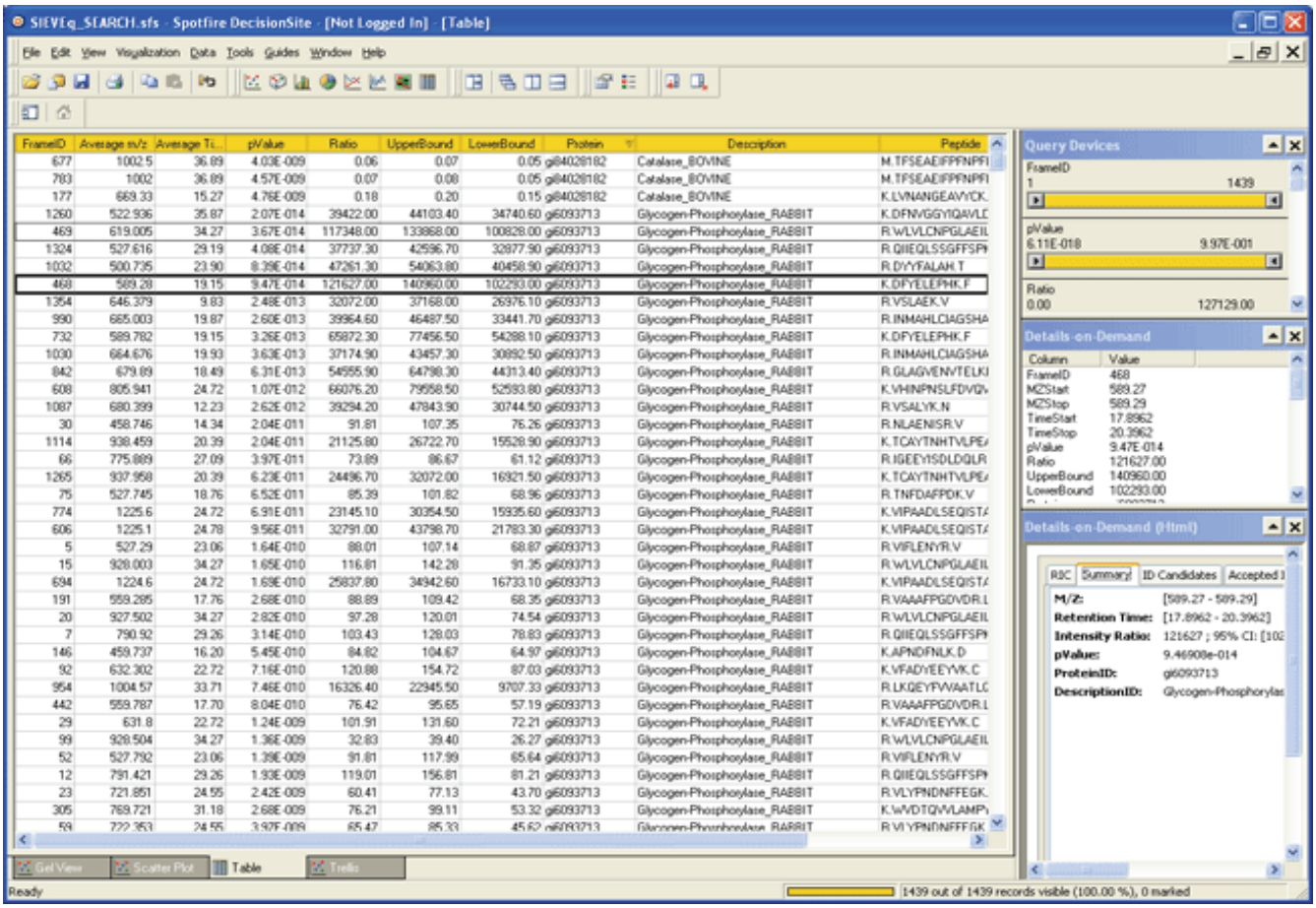


Table 4. Parameter definitions for the Results Summary table

Parameter	Definition
FrameID	A unique identifier for the frame. The number indicates the order in which the frame was triggered based on the MS peak intensity. For example, a FrameID of 3 indicates this was the 3rd most intense MS peak that triggered the creation of a frame.
Average *m/z*	Average *m/z* value. Calculated from the *m/z* start and *m/z* stop per frame.

Table 4. Parameter definitions for the Results Summary table, continued

Parameter	Definition
Average time	Average retention time, calculated from the time start and time stop per frame.
pValue	pValue of the t-test in order to provide a probability that the two distributions (the log intensities of the control and treatment) are different. A smaller pValue indicates a more statistically significant difference was measured.
Ratio	The average of the log peak intensities for each replicate in each group. LSample = Average(IPeak1, IPeak2,) |Sample, and LControl = Average(IPeak1, IPeak2,) |Control. The ratio is the ratio of these numbers (LSample / LControl).
UpperBound	The 95% confidence interval for this ratio based upon the statistics derived from the replicates. The 95% confidence interval is two numbers: a lower limit and an upper limit within which there is a 95% probability that the true ratio resides.
LowerBound	The 95% confidence interval for this ratio based upon the statistics derived from the replicates. The 95% confidence interval is two numbers: a lower limit and an upper limit within which there is a 95% probability that the true ratio resides.
Protein	An ID number for identified proteins.
Description	Description of the protein or the state of the frame if no protein is identified.
Peptide	Name of the identified peptide.
XCorr	The SEQUEST cross-correlation score of the top candidate peptide or protein. (Cross-correlation: the multiplication of two signals averaged over a time interval.)
Charge	The charge of the peptide identified from the database search.
MS2	The number of ms2 scans that belong to the frame. This number is dependent on the search width SIEVE parameter and is a sum from all the raw files used in the analysis.
Intensity_Control_[Control File number]	Highest intensity measurement from the specific raw file.
Intensity_Sample_[Sample File number]	Highest intensity measurement from the specific raw file.

Index

P

R

S

V